THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLV APRIL 1966 NUMBER 4

Design Considerations and Proposals for Compatible U. S. Subsidiary Coinage

By R. A. KIMBER and R. R. STOKES

(Manuscript received August 16, 1965)

For 173 years since the Coinage Act of 1792, United States subsidiary coinage has contained 90 per cent silver. During the last decade, an increasing unbalance between supply and demand has jeopardized the availability of silver and made it necessary to consider alternative coinage materials. Because present coin-handling mechanisms are designed to discriminate between coinage and spurious materials, it is necessary to identify the functional properties of coins so that a compatible material system may be proposed. Compatibility is established by considering such items as weight, resistivity, diameter, thickness, wear, scrap, composition, coinability, corrosion and color. One alloy and four laminated metal systems, varying in silver content from 70 to 0 per cent, were found which satisfy the compatibility requirements. This paper discusses the design considerations for compatibility and proposes five metal systems which meet these considerations.

I. INTRODUCTION

Coin silver, as used in U. S. subsidiary coinage, is made of an alloy containing 90 per cent silver and 10 per cent copper. During the last decade, it has become increasingly evident that the availability of this alloy for the continued production of coinage is jeopardized by an increasing unbalance between the demand for, and the supply of silver.

The involvement of Bell Telephone Laboratories in this issue results from the use of 5, 10, and 25 cent coins in public telephones. In view of the possible change in subsidiary coinage alloy, it became necessary to identify the functional properties of coins in their various applications and to specify accurately those properties which are essential to maintaining an operationally compatible coinage system. To this end, work was initiated on a study of coin properties and compatible metal systems for U. S. subsidiary coinage.

This paper records (*i*) the results of an investigation of the permissible deviations from the present coinage system for compatible performance in coin chutes, (*ii*) proposals for changes which represent improvements in the present coinage, and (*iii*) the analysis and description of various metal systems which could be employed. This information has been discussed and confirmed with various representatives of the vending industry and conveyed to the U. S. Mint and the Battelle Memorial Research Institute, consultant to the Mint, for their use.

II. COIN CONSIDERATIONS

2.1 *Functional Requirements*

The selection of alternative metal systems to provide compatible U. S. subsidiary coins involves a number of diverse considerations. There are practical questions relative to manufacture and procurement, psychological questions of public acceptance, and functional questions of serviceability in coin operated machines. It is the purpose of this paper to identify the functional requirements for subsidiary coins and to present specific criteria for insuring compatibility. This is facilitated by first discussing those factors which pertain to coin operated mechanisms.

2.1.1 *Weight*

The published weight limits on dimes and quarters as minted by the U. S. Department of the Treasury are 38.58 ± 1.5 grains (2.5 grams nominal) and 96.45 ± 3 grains (6.25 grams nominal), respectively. Coins in circulation naturally weigh less due to wear. To establish the reduction in size, a study was made of the physical properties of circulated coins which included data on changes in coin weight, diameter, and thickness. The sample included 1000 coins of each denomination, pro-rated to have representative numbers by date for the number of each remaining in circulation. Probability plots of the data showed that the population distribution was sufficiently normal to permit statistical estimation of the minimum and maximum values. The three-sigma limits, which theoretically include 99.73 per cent of the population, showed that the weight of dimes could range from a maximum minted weight of 2.6 grams to a low of 2.31 grams, while the weight of quarters could similarly range from 6.44 to 5.74 grams.

In coin equipment, the lightest coin must have sufficient energy to operate one or more mechanisms. Since the quarter is two and a half times heavier than the dime, the half dollar is five times heavier, and each is required to accomplish similar functions, it is safe to assume these requirements will be satisfied for the quarter and half dollar if they are satisfied for the dime. To establish the lower acceptable weight for a dime, coins of various weights were deposited in new and existing coin mechanisms. The data show that dimes weighing less than 1.98 grams begin to malfunction and cause misregistration of coin deposit information.

The wear study showed that a minted dime of 2.50 grams could wear to a lower expected weight of 2.31 grams, experiencing in circulation a maximum weight loss of 0.19 grams. Assuming the same wear characteristics for an alternate metal, the minimum minted weight of a new dime should not be less than 2.17 grams to avoid encountering worn dimes as light as 1.98 grams. The 2.17 gram dime weighs approximately 87 per cent of the present nominal minted value of 2.5 grams.

As previously mentioned, the weights of new quarters and half dollars are not critical for coin mechanism operation. Since most metals from which coins might be made have densities which differ from coin silver by less than 20 per cent, it is not considered restrictive to limit, arbitrarily, the weights of new quarters and half dollars to 80 per cent of their present weights. The reason for not reducing the weight further is to exclude the use of aluminum and magnesium in coins because of the inability to separate them by eddy current means. This will be discussed next.

2.1.2 *Acceptance Number*

Acceptance number is the nomenclature for a measure of a coin's sensitivity to eddy current detection. The lower the number, the greater the sensitivity, although the relationship is not linear.

An eddy current detector consists of a ramp on which a coin, or coin substitute, can roll through a magnetic field as shown in Fig. 1. When a coin rolls through the magnetic field, each incremental mass of the material instantaneously rotates about that point on the coin periphery which touches the ramp, and each cuts the transverse magnetic flux lines with some instantaneous velocity. By the "generator" rule, eddy currents flow in the coin perpendicular to the instantaneous velocity vectors. The lower the resistivity of the material, the greater are the eddy currents for a given velocity and magnet strength.

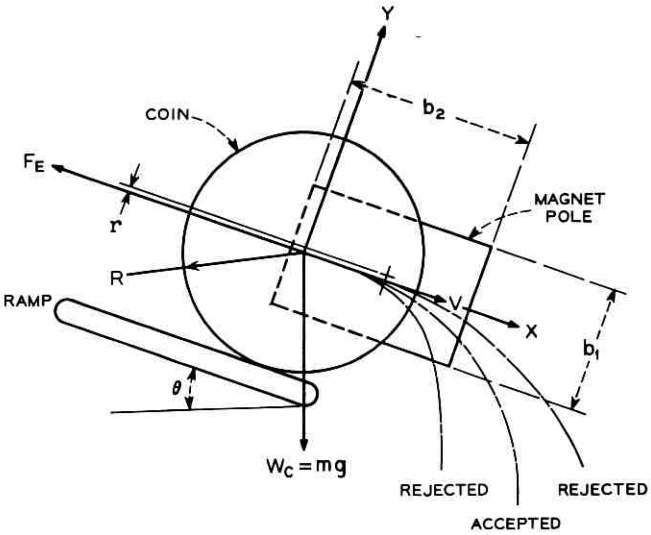


Fig. 1 — Relationship of coin, ramp, and magnet in an eddy current detector.

The induced currents flowing in a conducting material have a subsequent interaction with the magnetic field and give rise to mechanical forces as expressed by the "motor" rule. The geometry of the ramp, magnetic field, and rolling coin are such that the resultant of the instantaneous force vectors is in a direction opposite to the translational velocity of the coin and functions to oppose the velocity and retard the coin. The greater the eddy currents, the greater the retarding, or braking force, and the more the coin is slowed down. On the other hand, the greater the density of the coin the greater is its momentum and the more difficult it is to slow the coin down with a given retarding force.

Considering the joint effect of the coin material properties of resistivity and density on eddy current detection, a decrease in the value of either causes increased braking. The acceptance number of any material is defined as the product of its resistivity and density ($\sigma\rho$). Consequently, materials with low acceptance numbers experience more retardation and are said to be more sensitive to eddy current action.

The equation of motion for a homogeneous conductive disc rolling without slipping through a magnetic field is derived in Appendix A and is given below:

$$\frac{d^2x}{dt^2} + \frac{k\phi^2}{\sigma\rho}\cdot\frac{dx}{dt} - a = 0, \tag{1}$$

where

x = the linear progression of the center of the coin,
t = the length of time the coin experiences the magnetic field,
ϕ = total magnetic lines of flux,
k = a numerical constant,
σ = electrical resistivity,
ρ = density, and
a = the constant acceleration of a coin rolling down an incline.

A diagram showing the physical interrelationships of the members is given in Fig. 1. In eddy current terminology, the coefficient of the velocity term (dx/dt) is called the braking constant and is designated by K. Thus,

$$K = \frac{k\phi^2}{\sigma\rho}. \tag{2}$$

The coin material resistivity (σ) and density (ρ) appear in the denominator as a product ($\sigma\rho$) and is part of the rationale for so defining acceptance number. As this number decreases, the braking constant increases, consistent with the earlier observation that decreases in acceptance number result in increases in coin braking. Also, consistent with expectation, the braking constant increases with increasing magnet strength. Fortunately, the acceptance numbers of metals are sufficiently different to provide a basis for achieving discrimination between many of the common metals and U. S. coin silver. Table I lists a number of metals and their acceptance numbers. While the difference in acceptance numbers is not ideal, it is adequate, when used in conjunction with weighing, to permit a high degree of slug rejection.

To illustrate how physical separation actually occurs, reconsider the eddy current brake previously described as a ramp intersecting a transverse magnetic field (Fig. 1). If this ramp is terminated after braking occurs, so that coins can fall through free space, it is apparent that coins which have been slowed down more will assume vertical free fall closer to the end of the ramp. Coins which have been slowed down less will have more horizontal momentum and will travel further from the end of the ramp before assuming vertical free fall.

In experimental studies it was noted that the free-fall path of coin silver alloy is closer to aluminum than it is to zinc. This does not provide optimum spacing for obtaining separation from both metals. Since new metal systems are being proposed to replace coin silver alloy, it seems judicious to select one with a higher acceptance number to obtain opti-

TABLE I — ACCEPTANCE NUMBER OF TYPICAL METAL SYSTEMS

Material System	Electrical Resistivity (microhm-cm)	Density (grams/cm^3)	Accpt. No. (microhm gm/cm^2)
Aluminum	2.8	2.7	7.5
Magnesium	4.45	1.74	7.75
Aluminum base alloy	5.5	2.64	15.1
Silver, pure	1.60	10.5	16.8
Magnesium base alloy	10.0	1.78	17.8
U. S. coin silver alloy	2.1	10.3	21.6
40-58-2 silver-copper-zinc	2.3	9.5	21.8
70-27-3 silver-copper-zinc	2.25	9.8	22.1
97.5-2.5 copper-nickel	2.5	8.9	22.5
Zinc	5.92	7.13	42.2
70-30 brass	6.98	8.4	58.6
Nickel	6.84	8.9	60.8 (magnetic)
Ingot iron	9.7	7.76	75.2 (magnetic)
Grade A phosphor bronze	10.7	8.8	94.0
Columbium and alloys	13.1 min	8.6	122 min
Lead	20.6	11.2	231
18% nickel silver	29	8.7	252
Titanium	55	4.6	253
Zirconium	40 min	6.5	260 min
75-25 copper-nickel	32	8.9	285
95-5 nickel-silicon	38	8.55	325
18-8 stainless steel	79	8.0	632
90-10 nickel-chromium	80	8.7	696
Incoloy alloy 800	92	8.03	740
80-20 nickel-chromium	108	8.55	925
Nonconductors	∞	—	∞

mum rejection capabilities. Experimental studies indicate that a coin with an acceptance number of approximately 25 gives optimum rejection capabilities between aluminum and zinc. It is interesting to note that the Mercury dime had an acceptance number of approximately 25. With improvements in the refining of silver, the acceptance number of the Roosevelt dime has dropped to its present value of 21.6. This point is illustrated in Fig. 2, where the acceptance curve for the Mercury dime is closer to that of zinc than the Roosevelt dime. From the discussion to ensue on compatible coinage metals systems, it will become evident that the option to adopt a different acceptance number can be realized in a number of different proposals.

Even with the acceptance number of coin silver being closer to aluminum than zinc, the adequacy of eddy current coin chutes for separating spurious slug materials is readily demonstrated. A nationwide sample of over $\frac{1}{2}$ million nonstandard deposits in public coin operated equipment which did not have eddy current coin chutes was collected,

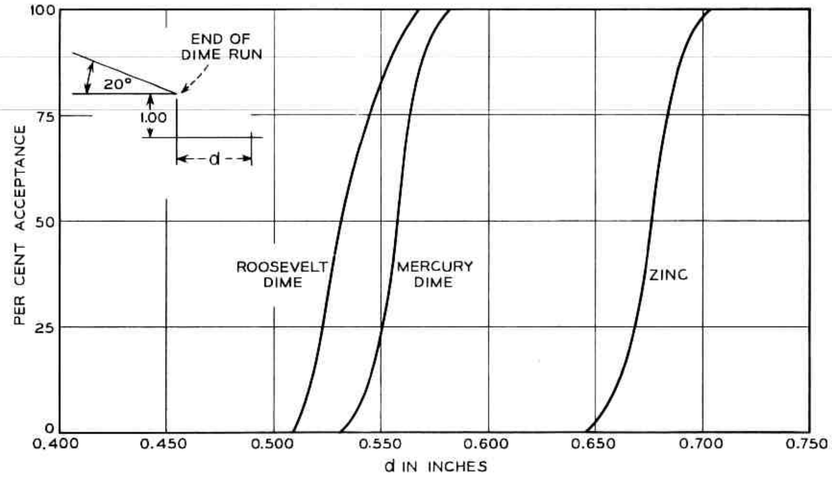


Fig. 2 — Acceptance vs coin position. (Eddy current magnetic field characteristics: $\phi = 1860$ lines; gap = pass 0.058 inch, stop 0.060 inch.)

sorted and identified. Of the total, approximately 73 per cent were unredeemable slugs and the remainder were tokens and foreign coins. Excluding the sample of tokens, foreign and mutilated coins which are either partly or totally redeemable, the composition of the slug sample is shown in Table II. The information is presented in categories representing the principal separation means in coin chutes. It is seen from this analysis that approximately 99.8 per cent of the sample would be rejected in eddy current coin chutes.

2.1.3. *Diameter*

The diameters of newly minted dimes and quarters are 0.705 ± 0.003 and 0.955 ± 0.003 inches, respectively. The previously mentioned coin study showed that the diameter of dimes in circulation ranged between 0.698 and 0.706 inches, while those for the quarter ranged between 0.946 and 0.957 inches. In many coin mechanisms, diameter gauging is the principal means of separation.

Further, discrimination against foreign coins with compatible alloys is achieved primarily on a diameter gauging basis. In view of these factors and the tremendous amount of coin handling, sorting, storing, and dispensing equipment which would have to be changed, it is not advisable to change the present minted diameters.

Table II — Material Composition Study of Nonstandard Deposits Collected in Public Coin-Operated Mechanisms Not Using Eddy Current Coin Chutes

		No. Nonstandard Pieces	Per Cent Of Total	
I. Would be rejected by eddy current coin chutes:				
Removed by magnetic trap:				
Iron	242,561			
Magnetic stainless steel	3,444			
	246,005	246,005	46.65	
Removed by weighing:				
Aluminum	2,311			
Plastic	1,705			
	4,016	4,016	0.76	
Removed by sizing:		92,041	17.46	
Removed by eddy current rejection:				
Lead	12,326			
Brass	99,788			
Zinc	13,522			
Nonmagnetic stainless steel (10¢ & 25¢ sizes)	14,011			
Tokens	5,906			
Other	38,628			
	184,181	184,181	34.93	
		526,243	99.80	99.80
II. Would be accepted in eddy current coin chutes:				
Nonmagnetic stainless steel	1,036		0.20	
		1,036		0.20
		527,279		100.00

2.1.4. *Thickness*

The thicknesses of newly minted dimes and quarters are 0.053 (+0.005, −0.003) and 0.067 (+0.005, −0.003) inches, respectively.[1] By the same criteria adopted for establishing minimum and maximum weights and diameters, the thicknesses of dimes in circulation range from 0.042 to 0.054 inches while those for quarters range from 0.053 to 0.072 inches.

In general, the thickness dimension is not a good measure of a coin. During use, mushrooming of the rim can cause increases in thicknesses;

in other cases, wear on the rim causes reduction in thicknesses. Further, many different foreign coins have substantially the same thicknesses. Thickness gauging is primarily useful in detecting mutilated coins, but is otherwise rather insensitive for regular gauging.

Although the coin gauge of the present coin mechanisms provides thickness gauging, the control is not so close that a small increase in thickness could not be tolerated. This may be necessary to help maintain the coin weight in a substitute metal system. Coin silver alloy has a density of 10.3 grams/cm^3 while most of the metals which might feature in a new system have lower densities. Two principal contenders, copper and nickel, each have nearly equal densities of approximately 8.9 grams/cm^3. Especially in the case of the dime, some change will have to be made to maintain the required weight.

2.1.5. *Wear*

Due to the nature of coin handling, coins wear primarily in thickness. Pearson product-moment correlations obtained from a sample coin study on coin wear indicates that there is a $+0.46$ correlation between coin weight and thickness, and practically no correlation between coin weight and diameter for coins in circulation. Because the wear occurs principally on the coined surfaces, it is apparent that a percentage change in coin thickness does not produce the same percentage change in coin weight as would be experienced with a solid disc. Over the range of expected coined surface wear, the loss of weight is estimated to be approximately one third of what would be experienced with a solid disc having the same thickness reduction. This was determined by comparing the coin weight losses with changes in coin thicknesses.

2.2 *Criteria for Compatibility*

From the previous section on factors pertaining to coin operated mechanisms, it is possible to identify a set of numerical constraints on coinage which forms the basis for an objective definition of compatibility. These constraints call for maintaining the present diameter, establishing a tolerance range for the acceptance number, and specifying values for the minimum weight and maximum thickness. The controlling physical dimensions for material and compatibility with U. S. dimes, quarters, and half-dollars are given in Table III.

2.3 *Laminated Coinage*

From the previous discussion on the relative uniqueness of the acceptance numbers of various metals, it is not surprising that the number

TABLE III — PHYSICAL CHARACTERISTICS FOR COMPATIBLE U. S. SUBSIDIARY COINAGE

	10¢	25¢	50¢
Diameter (inch)	0.705 ± 0.003	0.955 ± 0.003	1.205 ± 0.003
Acceptance number (microhm-gm/cm^2)	$25^{+1}_{-3.5}$	$25^{+1}_{-3.5}$	$25^{+1}_{-3.5}$
Weight (grams)	2.17 min	5.00 min	10.0 min
Thickness after coining (inch)	0.061 max	0.072 max	0.091 max

of existing alloys which satisfy the eddy current considerations is limited. When additional requirements like appearance, corrosion, and manufacturability, to be discussed later, are added, the number of possible choices is even more restricted. In view of this, attention has been directed to combining various metals in a laminate structure and utilizing the different material properties to achieve the equivalent acceptance number of coin silver alloy. One can get an intuitive understanding of the approach by considering a low resistivity metal, which permits high eddy currents and provides braking action to oppose the coin motion, laminated to a high density metal which provides inertia to assist the coin motion. By establishing the correct laminate thickness, it is possible to control the resulting acceptance number. This approach has the additional attraction of using common metals in an uncommon way to offer protection against counterfeiting.

The equation of motion for a laminated coin rolling without slipping through a magnetic field, given below, was derived and experimentally verified as presented in Appendix A.

$$\frac{d^2x}{dt^2} + \frac{k\phi^2 \sum_{i=1}^{n}\left(\frac{\tau_i}{\sigma_i}\right)}{\sum_{i=1}^{n}(\rho_i\tau_i)}\cdot\frac{dx}{dt} - a = 0, \tag{3}$$

where

τ_i = thickness of the ith lamella,
σ_i = resistivity of the ith lamella,
ρ_i = density of the ith lamella,
n = number of lamellae,

and all other terms are the same as defined in (2).

A comparison of this equation with that for the homogeneous coin shows that it differs only in the braking constant. Since the two equations have the same form, they have the same general solution. If the two braking constants were constrained to be equal, the equations would have the same specific solution and the velocities of the homogeneous and laminated coins leaving the magnetic field would be equal. This is the condition for compatible eddy current materials. When the two braking constants are equated and all common terms are cleared, the following equation results and provides the constraint relation on the section thicknesses (τ_i) of the laminated coin to achieve the equivalent acceptance number ($\sigma\rho$) of a homogeneous coin.

$$(\sigma\rho)_{\text{eff}} = \frac{\sum_{i=1}^{n} \rho_i \tau_i}{\sum_{i=1}^{n} \frac{\tau_i}{\sigma_i}}. \tag{4}$$

This equation has been verified experimentally and used in the design of the laminated coins proposed in a later section for compatible coin substitutes.

2.4 *Additional Coinage Considerations*

2.4.1. *Scrap*

One important consideration in the production of coins is the efficient utilization of scrap. Present punch press operations produce a material stock skeleton containing 30 to 35 per cent scrap. This is reclaimed by melting, casting, and rerolling. The same procedure can be employed for any single element or alloy substitute.

In the use of laminate strip stock, the scrap reclamation problem places two constraints on the choice of the metals for laminate sections. First, when laminated scrap is melted, all of the constituent metals in the various laminate sections appear in the resultant alloy. Because refining is too costly, it is necessary to be able to enrich the reclaimed stock by adding one or more of the constituent metals to produce an additional quantity of one of the required laminate alloys. Second, in the enriching process it is necessary that the reconstructed alloy be produced in quantities that can be completely used to prevent stockpiling.

These two constraints lead to two conclusions. First, the constituent elements used in one laminate section must appear in the second laminate metal which is planned for reconstruction. Second, the percentage of the reconstructed laminate used in the coin must exceed the percentage of scrap produced. In the simplest case, should one laminate be a pure metal, like zinc, the other laminate material must be a zinc bearing alloy and must constitute at least 35 per cent of the coin volume. The specific details of these constraints will be discussed for the proposed laminate metal systems in Section III.

2.4.2. *Composition*

From an eddy current view, the metallic elements must work into some metal system which achieves the required acceptance number and remains essentially non-magnetic. Should the proposed metal system require alloying, some of the most promising pure metals like silver and nickel cannot be used together except in very small quantities because of the tendency to crack during rolling. Lead is excluded in laminated structures as a high density replacement for silver because it fails to bond. The phase relationships of copper-nickel and copper-aluminum alloys cause the resistivities of both to rise radically above the resistivity of either pure metal and limit their utility.

In addition to the scrap problem previously discussed and the considerations of corrosion, coinability, and color, to follow, the materials should be in good supply at reasonable cost. Zinc, copper, and nickel are all attractive in this regard.

2.4.3. *Coinability*

Two of the principal factors which contribute to the appearance of a coin are the amount of relief and sharpness of the coined surfaces. By these criteria, American coins rank among the finest in the world, and there is strong interest in maintaining the present high standard. To achieve this objective, it is necessary to employ a metal system which has an adequately low initial hardness and which will not work-harden during coining by an amount which would prevent complete filling of the characters on the coined surfaces.

Presently, coin blanks have to be capable of having an initial hardness of 28 to 29 Rockwell B in the annealed condition and not rising above a hardness of 76 to 77 Rockwell B in the fully-coined condition. This places a double constraint on the selection of coin substitute metals.

Alloys which might be considered for laminated coins, such as stainless steel, silver-copper-nickel alloy, copper-zinc, and silver-copper-zinc alloys having more than 30 per cent zinc, are too hard in the annealed condition for coining. Other choices for coin substitutes which are coin-able like the silver-copper and copper-nickel alloys, change rapidly in hardness with the addition of certain trace elements used to control the alloy resistivity. Specifically, the addition of 0.1 per cent silver to copper-nickel alloys raises the hardness from 26 Rockwell B to about 60 Rockwell B and 1 per cent silver raises the hardness to 70 Rockwell B. Such factors are important with respect to both manufacturability and acceptability.

2.4.4. *Corrosion*

A further constraint is placed on the choice of materials by corrosion. Any metal system which is acceptable for subsidiary coinage must neither tarnish in a manner to change appreciably the appearance of the coin after minting, nor produce corrosive residues which could soil the property of a possessor, or become dislodged in coin mechanisms and eventually cause malfunctions.

In the case of laminates, additional care must be exercised to insure that the adjacent laminates are adequately close in the electromotive series to minimize rim corrosion. In the use of dissimilar materials, it is to be expected that some corrosion will occur. However, the effects of this at a low level are adequately offset by the effective polishing which coins experience in use.

2.4.5. *Color*

Since the coin silver alloy used in higher value subsidiary coinage has been "white," or "silvery," since the first coinage act in 1792, the American public is accustomed to white coins. How far a new metal system can depart from this standard is a subjective question.

Silver-copper alloys with decreasing silver content take on a progressively increasing pink tone which finally becomes a rich bronze color. A number of independent subjective evaluations set a lower silver content limit, concluding that at or above a 70 per cent silver content, the alloys look subjectively similar to the present coinage. The American public has also grown accustomed to the material appearance of the U. S. five cent coin which is 75-25 Cupro-Nickel. While this alloy has a different "whiteness," it is apparently acceptable for a non-silver standard. It is felt that any new metal system should retain white coined surfaces.

III. METAL SYSTEMS FOR COMPATIBLE COINAGE

This section includes five proposals for metal systems which provide eddy current compatibility with existing U. S. silver alloy coinage. One basic alloy and four laminate solutions are presented.

3.1 *Alloys*

In surveying the extensive number of existing alloys for possible compatible systems, the most severe requirement — correct acceptance number — was initially used. This criterion was applied to a listing of alloys presented in The Bureau of Standards Publication, *The Mechanical Properties of Metals and Alloys*. Alloys found which satisfy this requirement are:

(*i*) aluminum alloys
(*ii*) magnesium alloys
(*iii*) copper-zinc alloys with less than 5 per cent zinc
(*iv*) copper-nickel alloys with less than 3 per cent nickel
(*v*) silver-copper alloys
(*vi*) silver-copper-zinc alloys with less than 5 per cent zinc.

The copper-aluminum alloys are excluded here because resistivities of the alloys are high, yielding unacceptably high acceptance numbers.

There are other criteria which further reduce the possibilities for acceptable alloys. The aluminum and magnesium alloys do not have sufficient weight to dependably rotate the coin separating devices in existing coin operated mechanisms. The copper-zinc and copper-nickel alloys are coppery in color and tend to tarnish. Therefore, these alloys are eliminated from further consideration.

The silver-copper alloy meets the weight requirement for all percentage compositions. Fig. 3 shows the variations of resistivity-density, and acceptance number for different amounts of silver. The acceptance number is everywhere between 21.5 and 23 microhm-gm/cm^2 over the range of 26-74 to 94-6 silver-copper. As previously discussed, the color of the silver-copper alloys changes from a bright, shiny white as the silver content is reduced, requiring that all alloys having less than 70 per cent silver be eliminated because of the color requirement.

Although zinc is added as a whitener, the silver-copper-zinc alloys have much the same properties as silver-copper alloys. Since its corrosion properties are the same, the silver-copper-zinc series must contain at least 70 per cent silver to prevent corrosion. Therefore, the only compatible alloy system found is comprised of silver-copper alloys which contain more than 70 per cent silver.

3.2 *Laminates*

A laminated coin has a more complex set of constrains than an alloy coin as noted in the coin considerations previously reviewed. A large

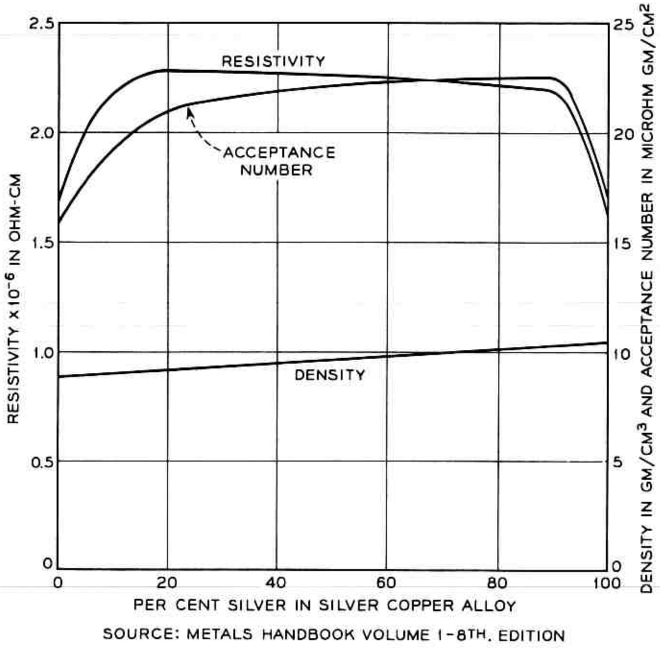


Fig. 3 — Resistivity and density of silver-copper alloys.

number of metal systems were investigated, and this section presents those few combinations which satisfy all of the restrictive conditions.

In the interest of achieving the simplest coin to fabricate, each laminated coin to be presented consists of three lamella sections. In each case, the two outside lamellae are white alloys to satisfy the color requirement, and are the same thickness to insure symmetrical eddy current performance. The core in each case is a copper alloy.

3.2.1 *70-30 Silver-copper sections laminated on a 30-70 silver-copper core.* (*Representative*)

In the discussion on alloys it was mentioned that all silver-copper alloys with silver contents between 26 and 94 per cent have acceptance numbers exceeding 21.5. The lower limit on silver content was set at 70 per cent for reasons of color and corrosion. This produced a coin material with an acceptance number of 22.5 which contained 70 per cent silver.

A laminated coin could be made with 70-30 silver-copper alloy on the outside layers for reasons previously cited, and reduced silver content alloy used in the core to achieve an over-all lower silver content in the coin.

The lower limit for the over-all silver content is determined by the minimum acceptance number. Assuming a dime coin blank thickness of 0.039 inch and minimum practical thicknesses of 0.007 inch for each outside laminate, the lowest silver content which can be used in the core and maintain an acceptance number of 21.5 is 20 per cent. This gives the coin an over-all silver content of 37.9 per cent.

Using this approach, it is possible to obtain a family of coins by letting the core contain higher silver than the 20 per cent minimum. For purposes of illustration, if a coin bearing 50 per cent silver were required, Fig. 4 would apply. To illustrate its use, if it is given that the outer laminates should be 70-30 silver-copper, the designer is free to select either the section thicknesses or the per cent silver in the core. For example, if the center core thickness is desired to be half of the total thickness, the core must contain 30 per cent silver. In this series of possible coins, there is no scrap problem since the melted scrap can be divided into two parts and each enriched to the desired value.

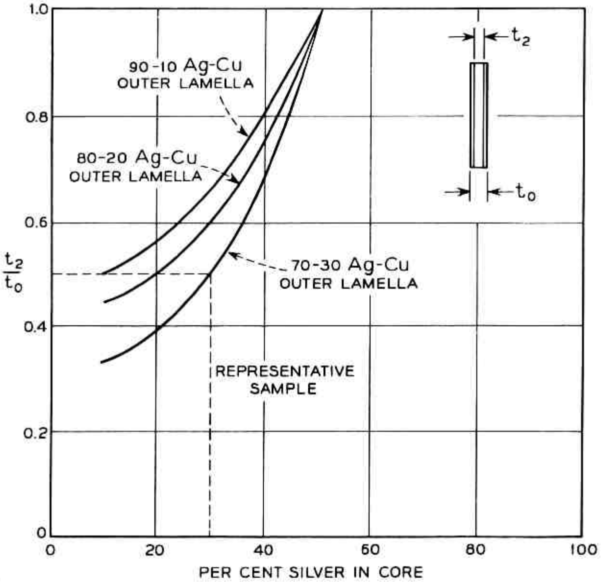


Fig. 4 — Laminated coin proportions. (Ag-Cu, Ag-Cu, Ag-Cu laminate. 50 per cent silver.)

3.2.2 *70-27-3 Silver-copper-zinc sections laminated on a 90-10 copper-zinc core*

The second laminated coin to be considered consists of two 70-27-3 silver-copper-zinc outer sections laminated on a 90-10 copper-zinc core and contains 44.5 per cent silver. The addition of zinc to copper in a ratio of one to nine increases the acceptance number to 34.5. The presence of zinc in the core complicates the reclamation problem by requiring that zinc also appear in the outer laminate material as discussed earlier in the section on scrap. Further, it needs to appear in the ratio of one to nine with the copper to simplify the melt-enriching process without causing stock piling. Alloys of 90-9-1 and 80-18-2 silver-copper-zinc were tried in addition to the proposed 70-27-3 composition. All work equally well from an eddy current view, but the 90 per cent silver alloy definitely leads to a scrap build up and the 80 per cent silver alloy situation was marginal. With the 70-27-3 alloy in the outer laminates and the 90-10 copper-zinc core, a coin is achieved which is completely compatible when the section thicknesses are properly controlled.

To visualize the effect of section thicknesses on the acceptance number, one might imagine the center thickness becoming increasingly thin, in which case the acceptance number would equal that of the outside material. Conversely, with increasing thickness of the inside section, the acceptance number approaches that of the center laminate. Depending on the thicknesses of the sections, the acceptance number of the coin could assume any value between the two limiting values of the constituent laminate alloys. To achieve the desired acceptance number of 25, it is obvious that one laminate material must have an acceptance number which is lower and the other material one that is higher. Obviously, as the outside laminates wear, the influence of the inner laminate increases and the composite acceptance number shifts towards that of the core material.

To determine what the section thicknesses should be to obtain a desired acceptance number, use is made of (4) which expresses the constraint relation. Since each coin consists of only two laminate materials, and since the cut blank thickness for a coin is constant (t_0), it is convenient to express the thickness of one laminate in terms of the other. In this case where the core material has the higher acceptance number, the inner laminate thickness is called t_2, and (4) takes the form

$$(\sigma\rho)_{\text{eff}} = \frac{[\rho_L(t_0 - t_2) + \rho_H t_2]}{\left[\frac{(t_0 - t_2)}{\sigma_L} + \frac{t_2}{\sigma_H}\right]}, \tag{5}$$

where

ρ_L = density of the lower acceptance number material,
ρ_H = density of the higher acceptance number material,
σ_L = resistivity of the lower acceptance number material,
σ_H = resistivity of the higher acceptance number material.

Since (5) is a function of only one independent variable, t_2, it is possible to plot a curve for the acceptance numbers which result from varying t_2. The resulting design chart is shown for this metal system in Fig. 5.

Since the cut blank thicknesses (t_0) for dimes, quarters, and half dollars are different, three separate design curves are shown for the three coin denominations and labeled $t_0 = 0.039$, $t_0 = 0.052$, and $t_0 = 0.066$, respectively. To obtain a coin acceptance number of 25, one has only to read the core material thicknesses, t_2, from the chart. The laminate thicknesses for the three denominations are given in Table IV.

Before leaving the design chart, it is appropriate to consider the change in acceptance number with wear. As a laminated coin wears, the outside sections are reduced in thickness while the core thickness is unchanged. This alters the ratio of the original thicknesses and causes a shift in the acceptance number. As the lower resistivity material on the out-

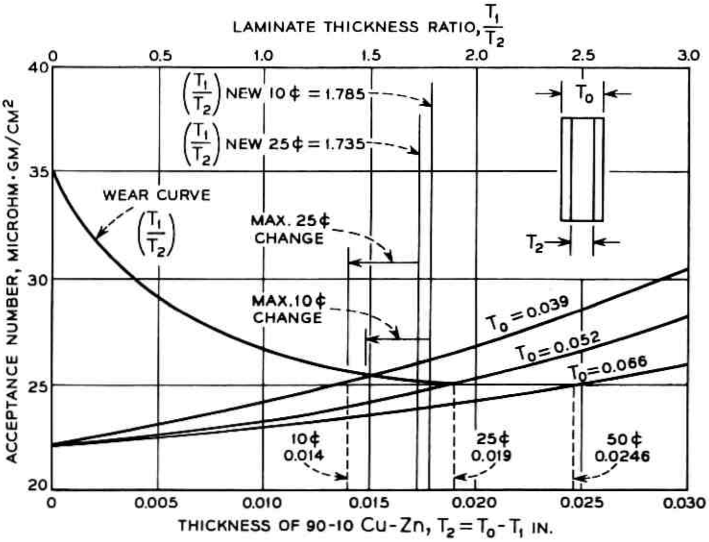


Fig. 5 — Laminated coin proposal no. 2. (70-27-3 Ag-Cu-Zn [$R = 2.25 \times 10^{-6}$ ohm-cm, $D = 9.8$ gm/cm^3] and 90-10 Cu-Zn [$R = 3.92$ ohm cm, $D = 8.8$ gm/cm^3].)

TABLE IV

Coin	70-27-3 Silver-Copper-Zinc	90-10 Copper-Zinc	70-27-3 Silver-Copper-Zinc
	$\frac{t_1}{2}$ (in)	t_2 (center) (in.)	$\frac{t_1}{2}$ (in.)
10¢	0.0125	0.014	0.0125
25¢	0.0165	0.019	0.0165
50¢	0.0207	0.0246	0.0207

side wears away, the acceptance number increases. Conversely, the acceptance number decreases if the denser material is used on the outside and wears away. To show this effect, a wear curve was constructed by holding the core thickness constant and calculating the acceptance number for reduced clad thicknesses.

Referring to the wear curve of Fig. 5, the ratio of combined outer laminate thicknesses to inner laminate thickness, t_1/t_2, is 1.785 for a newly minted 10-cent coin and this value on the wear curve coincides with an acceptance number of 25. From the study on coin wear, the maximum reduction in dime thickness was estimated to be 0.012 inches. Recalling that the weight decreased at one-third the rate of the thickness reduction, the maximum dime wear would result in the equivalent solid disc loss of 0.004 inch total, or 0.002 inch per side. Since this wear occurs on the two outer laminate sections, it reduces the combined outer section thicknesses t_1 to 0.021 inch, and the t_1/t_2 ratio to 1.50. Referring to the wear curve, this change would cause the acceptance number to rise to 25.2, which is an insignificant change. The wear limits for the dime and quarter acceptance number change differ as a function of the relative amount of wear and the initial thickness ratio. No wear curve is given for the 50-cent piece since wear data are not available.

As mentioned previously, the coins under discussion here contain approximately 44.5 per cent silver. The specific compositions of the three subsidiary coins are given in Table V.

The following calculation shows that there is no scrap build up with these coins. Assuming there is 100 pounds of laminate material before the coins are blanked, there will be 35 pounds of scrap to be melted. For the dime, this 35 pounds of scrap will consist of 15.75 pounds of silver, 17.32 pounds of copper, and 1.93 pounds of zinc. To achieve a new alloy in the same proportions as the outer lamella (70-27-3) only silver needs to be added to the melt, since the copper to zinc ratio re-

TABLE V — COIN COMPOSITION

	10¢	25¢	50¢
% Silver	45	44.4	44.0
% Copper	49.5	50.0	50.4
% Zinc	5.5	5.6	5.6
Total	100.0	100.0	100.0

mains nine to one. The weight of outer material obtained from the addition of silver is found from the following ratio

$$W_s = \frac{(\text{ \% Copper in melt}) \ (35 \text{ lbs scrap})}{\text{\% Copper in outer layer}} = \frac{17.32}{0.27} = 64.2 \text{ lbs.}$$

The weight of the outer lamella in 100 pounds of laminate is

$$W_0 = \frac{t_1}{t_2} \times 100 = \frac{0.025}{0.039} \ (100) = 64.2 \text{ lbs.}$$

Since the two weights are equal, there will be no scrap build up.

3.2.3 *40-50-5-5 Silver-copper-nickel-zinc sections laminated on a copper core*

The third proposal for a compatible metal system consists of two outer laminates of 40-50-5-5 silver-copper-nickel-zinc alloy bonded to a core of copper. The 40-50-5-5 silver-copper-nickel-zinc alloy is commonly known as Swedish Coin Silver. It is used in five Swedish coins, has an attractive white appearance and has fair coinability. As an alloy, the material has an acceptance number of 57.2 which is too high. As an outside lamella of a laminated coin with a copper core, it is an attractive material proposal for achieving an acceptance number of 25. The coin contains 19.9 per cent silver.

The design and wear curves for this coin were constructed as in the case of the previous coin, and are plotted in Fig. 6. In this case, since the material with the lower acceptance number is used in the core, the curve is plotted with the thickness, t_1, as the independent variable.

The coin meets all of the cited requirements and is attractive as a coin of low silver content. The section thickness for each denomination is given in Table VI with a listing of the various coin compositions.

3.2.4 *75-25 Copper-nickel sections laminated on a copper core*

This fourth proposal consists of two 75-25 copper-nickel outer sections laminated on a copper core and contains no silver. The 75-25 cupro-

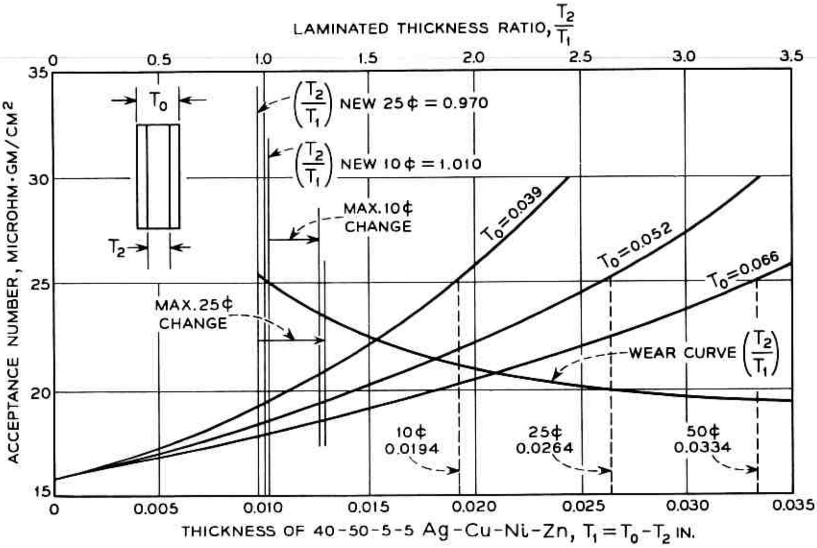


Fig. 6 — Laminated coin proposal no. 3. (40-50-5-5 Ag-Cu-Ni-Zn [$R = 6.05 \times 10^{-6}$ ohm-cm, $D = 9.45$ gm/cm^3] and Cu [R = 1.74×10^{-6} ohm-cm, D = 8.94 gm/cm^3].)

TABLE VI

Coin	40-50-5-5 Silver-Copper-Nickel-Zinc $\frac{t_1}{2}$ (in.)	Copper t_2 (center) (in.)	40-50-5-5 Silver-Copper-Nickel-Zinc $\frac{t_1}{2}$ (in.)
10¢	0.0097	0.0196	0.0097
25¢	0.0132	0.0256	0.0132
50¢	0.0167	0.0326	0.0167

COIN COMPOSITION

	10¢	25¢	50¢
% Silver	19.9	20.3	20.2
% Copper	75.1	74.6	74.8
% Nickel	2.5	2.55	2.5
% Zinc	2.5	2.55	2.5
Total	100.0	100.0	100.0

nickel is presently used in the coining of U. S. five cent pieces. It is an easy material to fabricate, is adequately white and has an acceptance number of 286. Because of this high value, it must be fabricated with thicker sections of a core material with a low acceptance number to achieve the desired acceptance number of 25. The acceptable composition percentages are given in Table VII.

The loss of outer laminate thickness due to wear has a more pronounced effect on the change in acceptance number than for the other laminates. Because of the faster relative change of the thickness ratio, the acceptance number is reduced after maximum wear to 22.5. This is not considered objectionable because the acceptance number remains within the proposed limits for compatible coinage. Experimental results with this coin composition confirmed the expectation that it yields better rejection than does present coin-silver.

IV. SUMMARY

After examining the properties of U. S. silver alloy coins and the operating requirements of typical coin handling mechanisms, it was concluded that the coin diameters should not be changed. The dime, quarter, and half-dollar minimum newly minted weights should be 2.17, 5.00, and 10.0 grams, respectively. The coined thicknesses of the quarter and half-dollar should be retained, and the dime thickness could be increased to 0.061 inch maximum to more easily achieve the weight objective.

Further, an acceptance number of 25 with a maximum and minimum limit of 26 and 21.5, respectively, would permit a distinct improvement in the ability to reject slugs.

It is possible to design a number of substitute metal systems which have eddy current compatibility with U. S. silver alloy coinage. After identifying the considerations which enter into the design of compatible coinage, five proposals have been made which satisfy these conditions. They differ in silver content from a high of 70 per cent to a low of zero, offering freedom in the final selection.

TABLE VII — COIN COMPOSITION

	10¢	25¢	50¢
% Nickel	9.9	10.1	10.2
% Copper	90.1	89.9	89.8
Total	100.0	100.0	100.0

Table VIII — Compatible Coin Proposals in Order of Merit

Order of Merit Metal System	Accpt. No. Before Wear	Accpt. No. After Max. Dime Wear	% Nominal Dime Weight Advantage Over Minimum
1. 70-27-3 Silver-Copper-Zinc Laminates on 90-10 Copper-Zinc	25.0	25.2	7.7%
2. 40-50-5-5 Silver-Copper-Nickel-Zinc Sections Laminated on a Copper Core	25.0	23.0	3.8%
3. 75-25 Copper-Nickel Sections Laminated on a Coppor Core	25.0	22.5	0.7%
4. Silver-Copper Alloys With at Least 70 Per Cent Silver	22.5	22.5	12.9%
5. 70-30 Silver-Copper Sections Laminated on a 30-70 Silver-Copper Core	22.1	21.6	9.7%

From a coin-operated mechanism point of view, there are three factors which provide a basis for comparing the five material systems: weight, acceptance number, and resistance to changes in acceptance number with wear. Because the minimum weight requirement is achieved with all five proposals, the ranking is based principally on the ability to obtain an acceptance number of 25 and to resist changes therein with wear, and secondarily, on the ability to obtain as much dime weight above the minimum as possible. Judging by these standards and using the dime data as a basis of comparison, it is possible to rank the five compatible coin proposals in an order of merit as listed in Table VIII.

V. ACKNOWLEDGMENTS

These proposals for compatible substitute coinage and the supporting technical information were submitted to Miss Eva Adams, Director of the U. S. Mint.

Appreciation is expressed to members of the U. S. Mint, the Battelle Memorial Institute, the Metals and Controls Division of Texas Instrument Company, and the Bell Telephone Laboratories' Metallurgical Engineering Department for their cooperative assistance.

APPENDIX A

Derivation of Eddy Current Equations

A.1 *Solid Disc*

The equation of motion for a homogeneous disc, rolling without slipping through a magnetic field, was derived by Messrs. L. Veith and C. F. Wiebusch using classical mechanics considerations and issued in an internal Bell Telephone Laboratories memorandum dated September, 1940. The essentials of these equations have, subsequently, been confirmed by derivations based on energy considerations and dimensional analysis. Because the latter technique affords a simple and direct approach to the equation of motion and emphasizes the important parameters from the onset, it is used in the following derivation. The authors are indebted to Mr. J. P. Runyon for suggesting its use in this paper.

Consider a homogeneous conductive disc rolling down a ramp without slipping and cutting a transverse magnetic field. The parameters, governing the motion of the disc, are defined below.

θ = angle of inclination of ramp.
x = linear progression of the center of the disc.
ϕ = magnetic lines of flux of the transverse magnetic field.
ρ = density of the disc.
σ = electrical resistivity of the disc.
a = initial acceleration of the disc.

The basic equation of motion for the dynamic system represented here is of the form

$$\frac{d^2x}{dt^2} + K_1 \frac{dx}{dt} + K_2 x + K_3 = 0. \tag{6}$$

An inspection of the equation shows that the constant, K_3, must have the dimensions of an acceleration, hence it is the initial acceleration, a, of the disc which is opposite to that due to the eddy current action. Further, since the disc is moving through a constant magnetic field, it is assumed that the motion is not a function of the displacement, x, and that the coefficient K_2 is zero. These considerations reduce the form of the equation to

$$\frac{d^2x}{dt^2} + K_1 \frac{dx}{dt} - a = 0. \tag{7}$$

Since the first and third terms have the dimensions of acceleration, $[LT^{-2}]$, where L is length and T is time, the second term must also have the same dimensions. Further, since the dx/dt part of the second term is a velocity with the dimensions of $[LT^{-1}]$, the coefficient K_1 must have the dimensions of $[T^{-1}]$. In eddy current terminology, the coefficient of the velocity term, K_1, is known as the braking constant.

From physical considerations, the braking constant is a function of the magnet and disc properties, specifically, the magnetic lines of flux, disc resistivity and density. This can be expressed in the form

$$K_1 \alpha \, \phi^{\delta} \rho^{\gamma} \sigma^{\lambda}. \tag{8}$$

Since ϕ has the dimensions of $[ML^2/TQ]$ where M represents mass units and Q represents charge units, ρ has the dimensions of $[M/L^3]$, σ has the dimensions of $[ML^3/TQ^2]$ and knowing that their product must have the dimension of $[T^{-1}]$, (8) can be written in the form

$$\left[\frac{1}{T}\right] = \left[\frac{ML^2}{TQ}\right]^{\delta}\left[\frac{M}{L^3}\right]^{\gamma}\left[\frac{ML^3}{TQ^2}\right]^{\lambda}. \tag{9}$$

Recombining

$$\left[\frac{1}{T}\right] = M^{(\delta+\gamma+\lambda)}T^{(-\delta-\lambda)}Q^{(-\delta-2\lambda)}L^{(2\delta-3\gamma+3\lambda)}. \tag{10}$$

Equating exponents on both sides of the equation leads to the following four equations:

$$-\delta - \lambda = -1 \tag{11}$$

$$\delta + \gamma + \lambda = 0 \tag{12}$$

$$-\delta - 2\lambda = 0 \tag{13}$$

$$2\delta - 3\gamma + 3\lambda = 0. \tag{14}$$

From the simultaneous solution of (11), (12), and (13),

$$\lambda = -1$$

$$\delta = 2$$

$$\gamma = -1.$$

Substitution in (8) yields the braking constant

$$K_1 \alpha \frac{\phi^2}{\sigma\rho}.$$

Therefore,

$$K_1 = k\frac{\phi^2}{\sigma\rho}, \tag{15}$$

and from (7) and (15), the equation of motion for a homogeneous disc rolling without slipping through a magnet field is

$$\frac{d^2x}{dt^2} + k\frac{\phi^2}{\sigma\rho}\frac{dx}{dt} - a = 0. \tag{16}$$

This equation assumes that the entire face area of the conductive disc is uniformly influenced by the magnetic field. In practice, this is rarely the case. Consequently, the braking constant is typically modified to account for the area and position of the disc relative to the area and position of the magnetic field. In addition, the edge effect of the disc entering and leaving the field must be considered. These factors are generally determined from geometric considerations and empirically obtained results in specific applications. In no event do these modifying coefficients reduce the generality of the basic equation of motion, given in (16).

A.2 *Laminated Disc*

The equation of motion for a laminated disc, rolling without slipping through a magnetic field, has the same basic form as that for a solid disc. An examination of (16) shows that all of the disc parameters appear in the braking constant. The specific form of the equation of motion for the laminated disc can be obtained by substituting appropriate expressions for the laminated disc resistivity and density in the braking constant. To facilitate this step, it is convenient to express the braking constant in terms of the disc mass and resistance.

The resistance of the conductive disc to eddy currents can be obtained from the general resistance equation

$$R = \frac{\sigma l}{A}. \tag{17}$$

With the coin center velocity parallel to the ramp and the magnetic field transverse to it, the direction of current flow is normal to the ramp along conductive length b_1, the height of the disc equal to the magnet dimension normal to the ramp. The cross-section area normal to the current is the product of the disc thickness, τ, and the width of disc equal to the length of the magnet, parallel to the ramp, b_2. Hence,

$$R = \frac{\sigma b_1}{\tau b_2}. \tag{18}$$

The density of the solid disc is simply ρ = mass/volume = $m/A_c\tau$. Substituting these expressions into the braking constant of (15) yields

$$K_1 = \frac{k\phi^2}{\left(\frac{R\tau b_2}{b_1}\right)\left(\frac{m}{A_c\tau}\right)} = \frac{k\phi^2 A_c b_1}{R m b_2}. \tag{19}$$

To obtain the appropriate expressions for the mass and resistance of a laminated coin, one can immediately write the mass equation

$$m_L = A_c(\rho_1\tau_1 + \rho_2\tau_2 + \cdots \rho_n\tau_n) = A_c\sum_{i=1}^{n}(\rho_i\tau_i) \tag{20}$$

where A_c = face area of the coin.

The total resistance of the laminated disc consists of the resistances of each lamination added as parallel resistors.

$$R_L = \frac{1}{\dfrac{1}{R_1} + \dfrac{1}{R_2} + \cdots + \dfrac{1}{R_n}} = \frac{1}{\sum\limits_{i=1}^{n} \dfrac{1}{R_i}} \tag{21}$$

where

$$R_i = \frac{\sigma_i b_1}{\tau_i b_2}$$

$$\therefore\ R_L = \frac{b_1}{b_2} \frac{1}{\sum\limits_{i=1}^{n} \dfrac{\tau_i}{\sigma_i}}. \tag{22}$$

Substituting in (19), the braking constant for a laminated disc takes the form

$$K_{1(\mathrm{eff})} = \frac{k\phi^2 \sum\limits_{i=1}^{n} \left(\dfrac{\tau_i}{\sigma_i}\right)}{\sum\limits_{i=1}^{n} (\rho_i \tau_i)}. \tag{23}$$

Hence, the equation of motion for the laminated disc is

$$\frac{d^2x}{dt^2} + \frac{k\phi^2 \sum\limits_{i=1}^{n} \left(\dfrac{\tau_i}{\sigma_i}\right)}{\sum\limits_{i=1}^{n} (\rho_i \tau_i)} \cdot \frac{dx}{dt} - a = 0. \tag{24}$$

A.3 *Equivalence of Solid and Laminated Discs*

Since the equations of motion for the homogeneous and laminated discs have the same form, they have the same solution and will differ in specific trajectories as a function of the braking constants only. The functional equivalence of the homogeneous and laminated discs can be obtained by equating the braking constants of (15) and (23), yielding

$$\frac{k\phi^2}{\sigma\rho} = \frac{k\phi^2 \left(\sum\limits_{i=1}^{n} \dfrac{\tau_i}{\sigma_i}\right)}{\sum\limits_{i=1}^{n} \rho_i \tau_i}. \tag{25}$$

Equation (25) reduces directly to the desired constraint relation.

$$\frac{1}{\sigma\rho} = \frac{\sum_{i=1}^{n} \frac{\tau_i}{\sigma_i}}{\sum_{i=1}^{n} \rho_i\tau_i}. \tag{26}$$

Equation (26) is in terms of the physical and material properties of the coins only, and establishes the equivalence between the two coins. While the resistivity-density product of a homogeneous disc is independent of thickness, it is readily seen that the value for the laminated coin is not. The expression correctly reduces to an identity for the case of a single laminate.

Putting (26) in a slightly different form,

$$(\sigma\rho)_{\text{eff}} = \frac{\sum_{i=1}^{n} \rho_i\tau_i}{\sum_{i=1}^{n} \frac{\tau_i}{\sigma_i}}, \tag{27}$$

it is apparent that the effective resistivity-density product for any laminated coin can be calculated with this equation.

A.4 *Experimental Verification of Equations*

To verify the equivalence of the laminar resistivity-density product, it was necessary to fabricate laminated coins using (27) and to test their dynamic performance against homogeneous coins of known parameters in an eddy current coin-sampling device. For this purpose, it was decided to build a laminar coin to simulate the resistivity-density product of zinc. Zinc was selected because it represents the next highest acceptance number above coin silver that the chute normally rejects.

The resistivity of zinc, determined with a Magnaflux Conductivity Meter (Model FM-100), is recorded in Table IX, along with the measured density. The acceptance number of this material is given as 42.3.

Table IX

Material	Resistivity × 10^{-6} ohm-cm	Density gm/cm^3	Acceptance Number $(\sigma\rho)$ × 10^{-6} gm-ohm/cm^2
Copper	1.74	8.86	15.4
Zinc	6.0	7.06	42.3
Phosphor Bronze (Grade A)	9.0	8.86	79.7

TABLE X

Material		Resistivity $\times 10^{-6}$ ohm cm		Density gm/cm³		Thickness inch		Acceptance Number ($\sigma\rho$) $\times 10^{-6}$ gm-ohm/cm²
1	2	σ_1	σ_2	ρ_1	ρ_2	t_1	t_2	
Phosphor Bronze	Copper	9.0	1.74	8.86	8.86	0.039	0.011	41.5

An examination of (27) indicates that the two materials selected to simulate a given coin must have acceptance numbers on each side of the desired value. For this reason, phosphor bronze (Grade A) with an acceptance number of 79.7 was laminated with copper to produce a zinc disc equivalent. The particular laminar thicknesses used in the simulated test discs are given in Table X, where t_2 is the thickness of the center laminate and t_1 is the combined thickness of the two, equal outside laminates. The tolerance on the three laminar section thicknesses was held to ±0.0005 inch each and the resulting calculated acceptance number of the clad disc was within 1.9 per cent of the homogeneous coin value. This deviation is considerably smaller than the discrimination capability of the eddy current test instrument. Therefore, the laminated discs were judged to be adequate test samples for comparative dynamic testing.

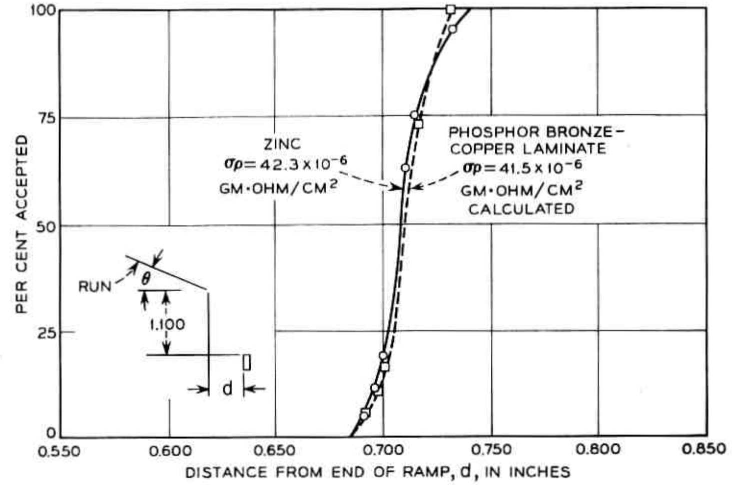


Fig. 7 — Acceptance vs coin position. (Magnet strength: ϕ = 1800 lines; θ = 20°; gap = pass 0.058 inch, stop 0.060 inch; coin diameter = 0.705 inch.)

It should be emphasized that the acceptance number of the laminated coin was calculated with the equation this experiment was designed to verify. The 1.9 per cent agreement, arrived at above, cannot be interpreted as a demonstration of equivalence. Satisfactory verification requires that the laminated coin exhibit the same dynamic behavior in the eddy current chute as the coin it was designed to simulate.

To evaluate the eddy current behavior of the laminated coins, a standard coin chute with the eddy current magnet adjusted to the nominal recommended flux strength of 1800 lines was mounted vertically. The accept mechanism was initially adjusted toward the front of the chute to insure that all deposited coins would be rejected. By progressively adjusting the mechanism toward the rear of the chute and dropping each coin and laminate equivalent 100 times at each setting, the results were obtained, as shown in Fig. 7, with the per cent acceptance plotted as a function of the distance, *d*, from the end of the run.

While there is some difference at the end points in the transition from low to high per cent acceptance, the mid-value acceptance characteristics of the solid coin and laminate disc equivalent approximate each other within the experimental accuracy of the test. This supports the conclusion that (27) is correct and provides an equivalence between single laminar and multiple laminate coins.

The end point transitional effect, referred to in the last paragraph, manifests itself as a sharper acceptance characteristic for the laminated disc. This means that the discrimination properties of this coin in eddy current detectors are superior to the homogeneous coin it simulates. Stated differently, the variation in trajectories which the laminated coin exhibits after leaving the run is less than for the companion homogeneous coin. This suggests that the laminated coin could be used not only to obtain a compatible eddy current coin alloy, but also, to achieve improved rejection characteristics.

REFERENCE

1. Annual Report of the Director of the Mint, Fiscal Year Ended June 30, 1962, U. S. Government Printing Office, Washington, 1963, p. 50.

Bounds on Communication with Polyphase Coding

By A. D. WYNER

(Manuscript received December 23, 1965)

The theoretical capabilities of a "polyphase" coding-modulation scheme with additive white Gaussian noise are studied. The channel capacity of this system is found and the error exponent estimated. Bounds are also found on $R_o(\rho_{\max})$, the maximum (asymptotic) rate for which polyphase codes can be found with maximum correlation between code words $\rho_{\max}$.

I. DEFINITIONS AND PRELIMINARIES

We shall consider the following ("polyphase") coding-modulation system (schematized in Fig. 1):

Every T seconds the message source emits one of M equally likely messages. The information rate is $R = 1/T \ln M$ nats per second. Corresponding to the ith message ($i = 1, 2, \cdots, M$) the coder emits an n-vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, where

$$-\pi \leqq x_{ik} \leqq \pi, \qquad k = 1, 2, \cdots, n, \tag{1}$$

and where the integer n will be specified later. The time interval $[0, T|$, during which this information must be transmitted, is divided into n equal subintervals of length T/n. During the kth of these subintervals, the modulated signal is

$$s_i(t) = \sqrt{2S} \cos (\omega_c t + x_{ik}), \qquad (k - 1) \frac{T}{n} \leqq t < \frac{kT}{n}, \qquad k = 1,2, \cdots, n. \tag{2}$$

Thus, we have employed phase modulation with carrier frequency ω_c radians per second and average power S.

We assume that the noise is additive, white, and Gaussian with one-sided spectral density N_o. The receiver must examine the received signal $y(t)$, the sum of $s_i(t)$ and the noise, and determine which of the M messages was actually transmitted. It is well known that (since all

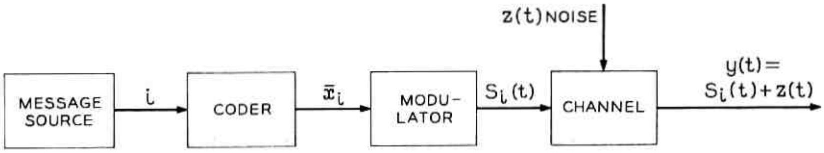


Fig. 1 — Polyphase coding-modulation system.

signals are equally likely to be transmitted, and have equal energy ST) the optimal decoder (which minimizes the average error probability) selects that signal $s_i(t)$ which maximizes ρ_i, the (normalized) correlation between $s_i(t)$ and $y(t)$:

$$\rho_i = \frac{1}{ST} \int_0^T s_i(t) y(t) dt. \tag{3}$$

Let us remark at this point that the correct operation of the decoder depends on its exact knowledge of the possible transmitted signals, so that in particular all delays and distortions to which the signals are subjected in transmission must be known exactly by the receiver. This is a so-called "coherent" receiver.

We let P_{ei} equal the probability that the decoder output is incorrect given that message i was transmitted, so that the average error probability is

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei} . \tag{4}$$

Now, the same channel is to be used by a number of users simultaneously, each at a different carrier frequency. Let W cycles per second be the separation of carrier frequencies between adjacent users (W will be taken as the "bandwidth"). Then the carrier frequency for the αth user (α an integer) is $\omega_c = \alpha 2\pi W$ radians per second. Further, we shall set $n = WT$ (let us say that T is such that WT is an integer), where n is the number of subintervals defined previously. With ω_c and n so chosen and the signals constructed as in (2), it is easy to show that the signals of the αth and βth ($\alpha \neq \beta$) users are orthogonal on the interval $[0, T]$. Hence, the presence of the signal due to the βth user does not affect the correlator in the decoder of the αth user.

Let us say that the transmission rate R and the bandwidth W are held fixed, and let T, the duration of the signals (hence $n = WT$), become large. Every T seconds the message source will produce one of $M = e^{RT}$ equally likely messages to which the coder must assign an n-vector. The channel capacity C is the maximum rate for which we may make P_e vanishing small for T sufficiently large. Formally, for any $R < C$ and $\varepsilon > 0$, there is a T sufficiently large so that the transmitter may transmit one of $M = e^{RT}$ messages with $P_e < \varepsilon$. (This will necessitate a set of $M = e^{RT}$ n-vectors stored in the coder.) The channel capacity C of this coding-modulation scheme is found in Section III.

Let us consider again the decoding scheme. Making use of the fact that the ρ_i of (3) are normally distributed random variables, it is possible to write an expression for the error probability P_e* which depends only on the signal energy to noise ratio ST/N_o and the matrix of normalized inner products among signals

$$\rho_{ij} = \frac{1}{ST}\int_0^T s_i(t)s_j(t)dt, \qquad i,j = 1,2, \cdots, M. \tag{5}$$

From (2) we obtain

$$\rho_{ij} = \frac{1}{n}\sum_{k=1}^{n} \cos\,(x_{ik} - x_{jk}), \qquad i,j = 1,2, \cdots, n. \tag{6}$$

It is known† that the error probability P_e (as given in (4)) using the optimal decoder may be bounded by

$$P_e \leqq f(\max_{i \neq j} \rho_{ij}),$$

where $f(x)$ is an increasing function of x. Accordingly, a reasonable procedure for designing good coding systems would be to try to make $\rho_{\max} = \max\limits_{i \neq j} \rho_{ij}$ as small as possible. Alternately we pose the problem as follows:

> With W, T, $\rho_{\max}$ held fixed, what is the largest rate for which we can design codes with parameters W, T, $\rho_{\max}$? (7)

Let us observe that from (6)

$$\begin{aligned}\rho_{ij} &= 1 - \frac{1}{n}\sum_{k=1}^{n} [1 - \cos\,(x_{ik} - x_{jk})] \\ &= 1 - \frac{1}{n}\sum_{k} 2\sin^2\frac{(x_{ik} - x_{jk})}{2} = 1 - \frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{2n}\end{aligned} \tag{8}$$

* Ref. 1, (2.11).

† Ref. 1, (4.7) and Ref. 2, p. 498.

where the "distance" $d(\mathbf{x}_i, \mathbf{x}_j)$ is defined by

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} \left[2 \sin \frac{(x_{ik} - x_{jk})}{2} \right]^2. \tag{9}$$

Thus, a code with maximum $\rho_{ij} = \rho_{\max}$, has minimum $d^2(\mathbf{x}_i, \mathbf{x}_j)/2n = (1 - \rho_{\max})$. In the light of the above, we shall reformulate the problem as follows:

Let $\mathcal{A}_n$ be the space of real n-vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ (where $n = WT$) which satisfy

$$-\pi \leqq x_k \leqq +\pi, \qquad k = 1, 2, \cdots, n. \tag{10}$$

Let $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n) \, \varepsilon \mathcal{A}_n$, and define the *distance* between $\mathbf{x}$ and $\mathbf{y}$ as

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{k=1}^{n} \left(2 \sin \frac{(x_k - y_k)}{2} \right)^2 \right]^{\frac{1}{2}}. \tag{11}$$

It will be shown in Section IV that $d(\mathbf{x},\mathbf{y})$ is, in fact, a metric. A *code* is a set of M members of $\mathcal{A}_n$, $\{\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})\}_{i=1}^{M}$. The *transmission rate* is $\hat{R} = 1/n \ln M$ nats per *symbol*. The transmission rate in nats per *second* is $R = (1/T) \ln M = W\hat{R}$. We will define $M(n,d)$ as the maximum number of code vectors in an n-dimensional code with minimum distance between pairs of code words d. Then $\hat{R}(n,d) = (1/n) \ln M(n,d)$, and $R(n,d) = (1/T) \ln M(n,d)$ are the corresponding transmission rates. A problem equivalent to that of (7) is the determination of $\hat{R}(n,d)$. In Section IV we shall let n (and hence T) become large while the ratio $\beta = d^2/2n$ is held fixed (corresponding to a fixed $\rho_{\max}$) and estimate $\hat{R}(\beta) = \lim_{n \to \infty} \hat{R}(n, \sqrt{2n\beta})$ by upper and lower bounds. Since $\beta = 1 - \rho_{\max}$, $\hat{R}(1 - \rho_{\max})$ is the (asymptotic) maximum rate for polyphase coding with $\max_{i \neq j} \rho_{ij} = \rho_{\max}$.

II. SUMMARY AND DISCUSSION OF RESULTS

The channel capacity is shown in Section III to be

$$C = W \left[-\int_0^\infty \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho} \, d\rho + \ln 2 \frac{A}{e} \right], \tag{12}$$

where

$$A = S/N_oW \tag{12a}$$

is the signal-to-noise ratio, and

$$\hat{f}(\rho) = 2\rho A e^{-A(1+\rho^2)} I_0(2\rho A), \tag{12b}$$

and $I_\nu(x)$ is the modified Bessel function of νth order. Another formula for C is (93). Approximate formulas for C for large and small values of the signal-to-noise ratio A are obtained in Appendix A. For large values of A,

$$C = \frac{W}{2} \ln \left(\frac{4\pi}{e} A \right) + \varepsilon_1(A), \tag{13}$$

where $\varepsilon_1(A) \to 0$ as $A \to \infty$. For values of A close to zero

$$C = W[A + 0(A^2)]. \tag{14}$$

The capacity C is plotted versus the signal-to-noise ratio A in Fig. 2. Estimates of the optimal achievable error probability are obtained in Appendix D.

The upper and lower bounds on $\hat{R}(\beta)$ are expressed in terms of the function $C_0(\xi)$ which is defined as follows. Let ξ be chosen

$$0 < \xi \leqq 1,$$

then define $\lambda(\xi)$ as the (unique) solution of

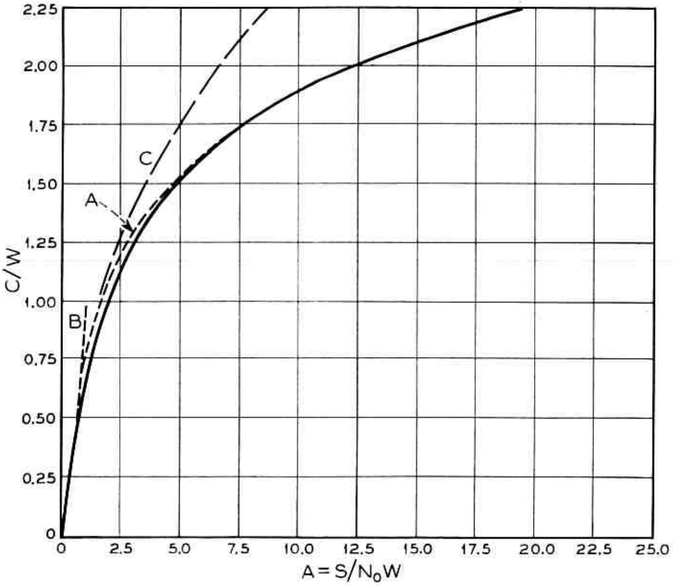


Fig. 2 — The channel capacity C vs the signal-to-noise ratio $A = S/N_oW$ — (12) (solid line). (Curves A and B are the approximations to the capacity C for large and small values of the signal-to-noise ratio A, respectively — (13) and (14). Curve C is $W \ln (1 + A)$, the capacity of a channel with bandwidth W and no restriction on the modulating scheme.

$$\xi = \left[1 - \frac{I_1(2\lambda(\xi))}{I_0(2\lambda(\xi))}\right]. \tag{15}$$

The existence (and uniqueness) of the solution to (15) is established in Appendix B. A graph of $\lambda(\xi)$ versus ξ is shown in Fig. 3. The function $C_0(\xi)$ is then defined as

$$C_0(\xi) = -\ln I_0(\lambda(\xi)) + (1 - \xi)\lambda(\xi). \tag{16}$$

A graph of $C_0(\xi)$ versus ξ is shown in Fig. 4. Our bounds on $\hat{R}(\beta)$, which are obtained in Section IV (and plotted in Fig. 5) are

$$C_0(\beta) \leqq \hat{R}(\beta) \leqq C_0(\gamma^2\beta), \tag{17}$$

where $C_0(\xi)$ is defined in (1) and

$$\gamma^2 = \frac{1}{\beta}(1 - \sqrt{1 - \beta}). \tag{18}$$

The lower bound is of the same type as the Gilbert bound for binary coding, and the upper bound makes use of the Blichfeldt density method.[3] Let us remark that the upper and lower bounds of (14) agree when $\beta = 1$, yielding $\hat{R}(\beta) = 0$ for $\beta \geqq 1$. When β is small it is shown in Appendix E that

$$\tfrac{1}{2}\ln\frac{1}{\pi e\beta} + \varepsilon_2(\beta) = \hat{R}(\beta) = \tfrac{1}{2}\ln\frac{2}{\pi e\beta} + \varepsilon_2(\beta), \tag{19}$$

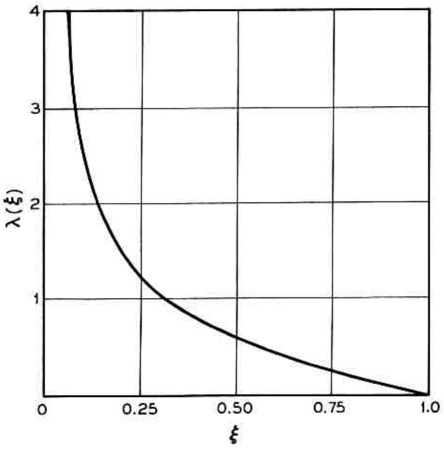


Fig. 3 — The function $\lambda(\xi)$ vs ξ — (15).

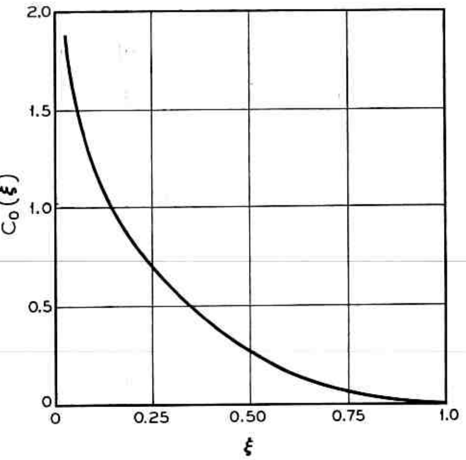


Fig. 4 — The function $C_0(\xi)$ vs ξ — (16).

where ε_1, $\varepsilon_2 \rightarrow 0$ as $\beta \rightarrow 0$. Thus, for sufficiently small β, $\hat{R}(\beta)$ is within $\frac{1}{2}$ ln 2 of $\frac{1}{2}$ ln $(2/\pi e \beta)$.

In terms of the modulation scheme discussed in Section I it is more revealing to rewrite inequalities (17) in terms of $\rho_{\max}$ the maximum correlation between pairs of signals. Let $R_0(\rho_{\max}, W, T) = R_0(\rho_{\max}, T)$ be the maximum rate (in nats per second) attainable for the polyphase

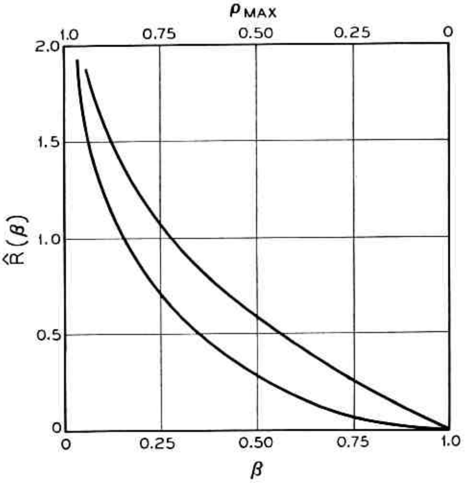


Fig. 5 — The upper and lower bounds $\hat{R}$ (β) vs β and $\rho_{\max} - 1 = \beta$ — (17). $\hat{R}$ (β) lies in the shaded region.

modulation scheme of Section I with parameters $\rho_{\max}$, W, and T. Let $R_o(\rho_{\max}) = \lim_{T\to\infty} R_o(\rho_{\max}, T)$. In the light of comment following (9), $R_o(\rho_{\max}) = R[(1 - \rho_{\max})]$. The upper and lower bounds on $R_o(\rho_{\max})$ are plotted versus $\rho_{\max}$ in Fig. 5.

Appendix F contains a comparison of the capabilities of this polyphase system and another important modulation system.

III. CHANNEL CAPACITY

The signals $s_i(t)$, $i = 1, 2, \cdots, M$, are of the form

$$s_i(t) = \sqrt{2S} \cos (\alpha 2\pi W t + x_{ik}), \qquad (k-1)\frac{T}{n} \leqq t < \frac{kT}{n}, \quad k = 1,2, \cdots, n, \tag{20a}$$

where $n = WT$ and

$$-\pi \leqq x_{ik} \leqq \pi, \qquad k = 1, 2, \cdots, n. \tag{20b}$$

Alternately, we may write

$$s_i(t) = x_{ik}^{(1)} \cos \alpha 2\pi W t + x_{ik}^{(2)} \sin \alpha 2\pi W t, \qquad (k-1)\frac{T}{n} \leqq t < \frac{kT}{n}, \quad k = 1,2, \cdots, n, \tag{21a}$$

where

$$x_{ik}^{(1)} = \sqrt{2S} \cos x_{ik}, \qquad x_{ik}^{(2)} = \sqrt{2S} \sin x_{ik}. \tag{21b}$$

The noise function $z(t)$ is a sample from a white Gaussian noise process with one sided spectral density N_o (so that the covariance is $R(\tau) = (N_o/2)\delta(\tau)$). The received signal is $y(t) = s_i(t) + z(t)$, where $s_i(t)$ is one of the M signals. The optimal decoder computes

$$\rho_i = \frac{1}{ST} \int_0^T s_i(t) y(t) dt,$$

and decodes $y(t)$ as that $s_i(t)$ with largest ρ_i. If $y(t)$ is the received signal, let $y^*(t)$ be

$$y^*(t) = y_k^{(1)} \cos \alpha 2\pi W t + y_k^{(2)} \sin \alpha 2\pi W t, \qquad (k-1)\frac{T}{n} \leqq t < \frac{kT}{n}, \quad k = 1,2, \cdots, n, \tag{22a}$$

where

$$y_k^{(1)} = 2W \int_{(k-1)(T/n)}^{kT/n} y(t) \cos \alpha 2\pi W t \, dt, \tag{22b}$$

and

$$y_k^{(2)} = 2W \int_{(k-1)(T/n)}^{kT/n} y(t) \sin \alpha 2\pi W t \, dt. \tag{22c}$$

We may think of $y^*(t)$ as the projection of $y(t)$ onto the space of allowable signals. It follows by direct computation that

$$\frac{1}{ST} \int_0^T s_i(t) y^*(t) dt,$$

the correlation of $y^*(t)$ and the ith signal $s_i(t)$, equals ρ_i . Thus, without loss of generality, we may consider the received signal to be $y^*(t)$. From (21) and (22), it suffices to consider the noise to be

$$\begin{gathered} z^*(t) = y^*(t) - s_i(t) = z_k^{(1)} \cos \alpha 2\pi W t + z_k^{(2)} \sin \alpha 2\pi W t \\ (k - 1) \frac{T}{n} \leqq t < \frac{kT}{n}, \qquad k = 1,2, \cdots, n, \end{gathered} \tag{23a}$$

where

$$\begin{gathered} z_k^{(1)} = y_k^{(1)} - x_{ik}^{(1)} \quad \text{and} \quad z_k^{(2)} = y_k^{(2)} - x_{ik}^{(2)}, \\ k = 1, 2, \cdots, n. \end{gathered} \tag{23b}$$

From (23b), (22b), (21b), and (20a) we may write

$$\begin{aligned} z_k^{(1)} &= 2W \int_{(k-1)(T/n)}^{kT/n} (y(t) - s_i(t)) \cos \alpha 2\pi W t \, dt \\ &= 2W \int_{(k-1)(T/n)}^{kT/n} z(t) \cos \alpha 2\pi W t \, dt, \qquad k = 1,2, \cdots, n, \end{aligned} \tag{24}$$

so that $z_k^{(1)}$ is a normally distributed random variable with mean zero and variance

$$\begin{aligned} E(z_k^{(1)^2}) = 4W^2 &\int_{(k-1)(T/n)}^{kT/n} \\ &\cdot \int_{(k-1)(T/n)}^{kT/n} \cos \alpha 2\pi W t \, (\cos \alpha 2\pi W \tau) \overline{z(t) z(\tau)} \, dt \, d\tau, \end{aligned} \tag{25}$$

where the over-bar denotes expectation. Since $\overline{z(t)z(\tau)} = R(t - \tau) = (N_o/2)\delta(t - \tau)$, the variance of $z_k^{(1)}$ is $N_o W$. Similarly for $z_k^{(2)}$. Further, $E(z_k^{(1)} z_k^{(2)}) = 0$, and

$$E(z_{k_1}^{(i)}, z_{k_2}^{(j)}) = 0 \qquad (i, j = 1, 2) \quad \text{if} \quad k_1 \neq k_2 .$$

Thus, these random variables are independent.

We conclude from the above that our channel is equivalent to the following *time-discrete memoryless* channel. Every $T/n = 1/W$ seconds, the channel input is a real number $X\varepsilon[-\pi,\pi]$. The output is a pair of numbers Y_1 and Y_2 given by

$$Y_1 = X_1 + Z_1 , \qquad Y_2 = X_2 + Z_2 \tag{26a}$$

where

$$X_1 = \sqrt{2S} \cos X, \qquad X_2 = \sqrt{2S} \sin X, \tag{26b}$$

and Z_1 , Z_2 are independent normally distributed random variables with mean zero and variance $N = N_0W$. Consequently, known results for determining capacity may be used.

If an input probability distribution is specified, the mutual information of the input and the output is

$$I(Y_1 , Y_2 ; X) = H(Y_1 , Y_2) - H(Y_1 , Y_2 \mid X), \tag{27}$$

where $H(Y_1 , Y_2)$ is the joint uncertainty of Y_1 and Y_2 and

$$H(Y_1 , \; Y_2 \mid X)$$

is the conditional uncertainty of Y_1 , Y_2 given X. The channel capacity C, in nats per *second* is

$$C = W[\max I(Y_1 , Y_2 ; X)], \tag{28}$$

where the maximization is performed over all possible input distributions. We proceed to find C.

Say $X = x$, and let $x_1 = \sqrt{2S} \cos x$, $x_2 = \sqrt{2S} \sin x$, then

$$H(Y_1 , Y_2 \mid X = x)$$
$$= \int_{-\infty}^{+\infty} \cdot \int_{-\infty}^{+\infty} dy_1dy_2g(y_1 - x_1 , y_2 - x_2) \ln g(y_1 - x_1 , y_2 - x_2), \tag{29a}$$

where

$$g(z_1 , z_2) = \frac{1}{2\pi N} \exp{[-(z_1^2 + z_2^2)/2N]} \tag{29b}$$

is the joint probability density of Z_1 , Z_2 . After changing the variables of integration and integrating (29a), we obtain,

$$H(Y_1 , Y_2 \mid X = x) = \ln 2\pi eN, \tag{30}$$

independent of x. Thus,

$$H(Y_1Y_2 \mid X) = \ln 2\pi eN, \tag{31}$$

independent of the input distribution.

Thus, to find C, we must maximize $H(Y_1, Y_2)$. Say $p_0(x)$ is the probability density of the input X, and $p_{12}(y_1, y_2)$ the resulting joint probability density of the output pair (Y_1, Y_2). If we characterize the output pair by polar coordinates $(\mathcal{R}, \Phi)$, then the corresponding density for $\mathcal{R}, \Phi$ is

$$f_{12}(r,\varphi) = rp_{12}(r \cos \varphi, r \sin \varphi), \qquad r \geqq 0, \qquad -\pi \leqq \varphi \leqq \pi, \tag{32}$$

where r is the Jacobian of the transformation. Hence,

$$\begin{aligned} H(Y_1Y_2) &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p_{12}(y_1, y_2) \ln p_{12}(y_1, y_2)dy_1dy_2 \\ &= -\int_{r=0}^{\infty}\int_{-\pi}^{\pi} p_{12}(r \cos \varphi, r \sin \varphi) \\ &\qquad \cdot \ln [p_{12}(r \cos \varphi, r \sin \varphi)]r\, dr\, d\varphi \\ &= -\int_0^{\infty}\int_{-\pi}^{\pi} f_{12}(r,\varphi) \ln \frac{f_{12}(r,\varphi)}{r}\, dr\, d\varphi \\ &= -\int_0^{\infty}\int_{-\pi}^{\pi} f_{12}(r,\varphi) \ln f_{12}(r,\varphi)dr\, d\varphi \\ &\qquad + \int_0^{\infty}\int_{-\pi}^{\pi} f_{12}(r,\varphi) \ln r\, dr\, d\varphi \\ &= H(\mathcal{R},\Phi) + \int_0^{\infty} f_1(r) \ln r\, dr, \end{aligned} \tag{33}$$

where $H(\mathcal{R}, \Phi)$ is the joint uncertainty of $\mathcal{R}, \Phi$, and $f_1(r)$ is the marginal density of $\mathcal{R}$. Now

$$H(\mathcal{R}, \Phi) \leqq H(\mathcal{R}) + H(\Phi), \tag{34}$$

(where $H(\mathcal{R})$, $H(\Phi)$ are the uncertainties of $\mathcal{R}, \Phi$, respectively) with equality if and only if $\mathcal{R}, \Phi$ are independent, and

$$H(\Phi) \leqq \ln 2\pi, \tag{35}$$

with equality if and only if Φ is uniformly distributed on the interval $[-\pi,\pi]$. Hence, from (33), (34), and (35),

$$H(Y_1, Y_2) \leqq H(\mathcal{R}) + \ln 2\pi + \int_0^{\infty} f_1(r) \ln r\, dr. \tag{36}$$

We shall now find $f_1(r)$, the density of $\mathcal{R}$, and show that it is independent of the input density $p_0(x)$. To begin with, let us say that $X = x$. Then the joint density of (Y_1, Y_2), given that $X = x$ is

$$\begin{aligned} &p_{12}(y_1, y_2 \mid X = x) \\ &= \frac{1}{2\pi N} \exp\{-[(y_1 - \sqrt{2S}\cos x)^2 + (y_2 - \sqrt{2S}\sin x)^2]/2N\}. \end{aligned} \tag{37}$$

The joint density of Y_1, Y_2 or the corresponding joint density of R, Φ is obtained from (37) by averaging over x:

$$\begin{aligned} f_{12}(r,\varphi) &= rp_{12}(r\cos\varphi, r\sin\varphi) \\ &= r\int_{-\pi}^{\pi} p_0(x)p_{12}(r\cos\varphi, r\sin\varphi \mid X = x)dx \\ &= r\int_{-\pi}^{\pi} p_0(x)dx\,\frac{1}{2\pi N} \\ &\quad\cdot\exp\left\{-\frac{1}{2N}[r\cos\varphi - \sqrt{2S}\cos x)^2 \right. \\ &\quad\left. + (r\sin\varphi - \sqrt{2S}\sin x)^2]\right\} \\ &= \frac{1}{2\pi N} re^{-(r^2+2S)/2N}\int_{-\pi}^{\pi} p_0(x) \\ &\quad\cdot\exp\left(\frac{r\sqrt{2S}}{N}\cos(x-\varphi)\right)dx. \end{aligned} \tag{38}$$

Now, the marginal density for R is obtained by integrating φ out of (38)

$$\begin{aligned} f_1(r) &= \int_{-\pi}^{\pi} f_{12}(r,\varphi)d\varphi \\ &= \frac{re^{-(r^2+2S)/2N}}{2\pi N}\int_{-\pi}^{\pi} d\varphi\int_{-\pi}^{\pi} dx\, p_0(x)\exp\left(\frac{r\sqrt{2S}\cos(x-\varphi)}{N}\right). \end{aligned} \tag{39}$$

Interchanging the order of integration, we get

$$\begin{aligned} f_1(r) &= \frac{re^{-(r^2+2S)/2N}}{2\pi N}\int_{-\pi}^{\pi} p_0(x)dx\int_{-\pi}^{\pi} d\varphi\exp\frac{r\sqrt{2S}}{N}\cos(\varphi - x) \\ &= \frac{re^{-(r^2+2S)/2N}}{N} I_0\left(\frac{\sqrt{2S}r}{N}\right), \end{aligned} \tag{40}$$

independent of $p_0(x)$.* We conclude from (40) and (36) that

$$\max_{p_0(x)} H(Y_1, Y_2) \leqq -\int_0^\infty f_1(r) \ln \frac{f_1(r)}{r}\, dr + \ln 2\pi, \tag{41}$$

where $f_1(r)$ is given by (40).

Let us now say that the input distribution is $p_0(x) = 1/2\pi$. Then from (38)

$$f_{12}(r,\varphi) = f_1(r)\, \frac{1}{2\pi}, \tag{42}$$

so that $\mathcal{R}$, Φ are independent with the marginal density of Φ, $f_2(\varphi) = 1/2\pi$. Thus, in this case, the equalities in (34) and (35) and hence in (36) hold yielding

$$H(Y_1, Y_2) = -\int_0^\infty f_1(r) \ln \frac{f_1(r)}{r}\, dr + \ln (2\pi), \tag{43}$$

so that (41) is satisfied with equality. From (28), (30), (41), and (43), the channel capacity C is given by

$$\frac{C}{W} = -\int_0^\infty f_1(r) \ln \frac{f_1(r)}{r}\, dr - \ln eN. \tag{44}$$

If we set $\rho = r/\sqrt{2S}$ and $A = S/N = S/N_oW$, the "signal-to-noise ratio", we obtain

$$\frac{C}{W} = -\int_0^\infty \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho + \ln \frac{2A}{e}, \tag{45a}$$

where

$$\hat{f}(\rho) = 2A\rho e^{-A(\rho^2+1)} I_0(2A\rho). \tag{45b}$$

IV. BOUNDS ON $\hat{R}(\beta)$

4.1 *Upper Bound on* $\hat{R}(\beta)$

We need the following two lemmas:

* We shall make frequent use of the formula

$$I_0(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{x \cos \theta}\, d\theta,$$

which can be found in Ref. 4, p. 79.

Lemma 1: Let $g_1, g_2, \cdots, g_p$ be real numbers. Then

$$\sum_{k=1}^{p} g_k^2 \geqq \frac{1}{p} (\Sigma g_k)^2. \tag{46}$$

Proof: From the Schwarz inequality

$$\left(\sum_{k=1}^{p} 1 \cdot g_k\right)^2 \leqq \left(\sum_{k=1}^{p} 1^2\right)\left(\sum_{k=1}^{p} g_k^2\right) = p \sum_{k=1}^{p} g_k^2. \tag{47}$$

Lemma 2: Let $\{\mathbf{x}_i\}_{i=1}^m$ be a set of m n-vectors from $\mathcal{A}_n$ with minimum distance d between pairs of vectors. The distance is given by (11). Let $\mathbf{y}$ be an arbitrary vector in $\mathcal{A}_n$, and denote by d_i the distance $d(\mathbf{x}_i, \mathbf{y})$. Then

$$\left(\sum_{i=1}^{m} \frac{d_i^2}{n}\right) - 4m\left(\sum_{i=1}^{m} \frac{d_i^2}{n}\right) + 2(m)(m-1)\frac{d^2}{n} \leqq 0. \tag{48}$$

Proof: Let us define a mapping of $\mathcal{A}_n$ into E_{2n}, Euclidean $2n$-space, as follows. If $\mathbf{x} = (x_1, x_2, \cdots, x_n) \varepsilon \mathcal{A}_n$, then the corresponding $2n$-vector is $\mathbf{x}' = (u_1, v_1, u_2, v_2, \cdots, u_n, v_n)$ where

$$u_k = \cos x_k, \qquad v_k = \sin x_k, \qquad k = 1, 2, \cdots, n. \tag{49}$$

Then letting $\mathbf{x}_1, \mathbf{x}_2 \varepsilon \mathcal{A}_n$, and letting $\mathbf{x}_1' = (u_{11}, v_{11}, u_{12}, v_{12}, \cdots, u_{1n}, v_{1n})$ and $\mathbf{x}_2' = (u_{21}, v_{21}, u_{22}, v_{22}, \cdots, u_{2n}, v_{2n})$ be the corresponding members of E_{2n}, the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\begin{aligned} d^2(\mathbf{x}_1, \mathbf{x}_2) &= \sum_{k=1}^{n} \left[2 \sin \frac{(x_{1k} - x_{2k})}{2}\right]^2 \\ &= \sum_{k=1}^{n} \{(u_{1k} - u_{2k})^2 + (v_{1k} - v_{2k})^2\}. \end{aligned} \tag{50}$$

To see this we need only observe that if the $x_{1k}, x_{2k}, k = 1, 2, \cdots, n$ are considered as arc lengths on a unit circle with center at the origin, then (u_{1k}, v_{1k}) and (u_{2k}, v_{2k}) are the Cartesian coordinates of x_{1k}, and x_{2k}, respectively, (see Fig. 6). The quantity $2 \sin [(x_{1k} - x_{2k})/2]$ is then the Euclidean distance between (u_{1k}, v_{1k}) and (u_{2k}, v_{2k}). Hence, $d(\mathbf{x}_1, \mathbf{x}_2)$ is the Euclidean distance between $\mathbf{x}_1'$ and $\mathbf{x}_2'$. This also provides a justification for calling $d(\mathbf{x},\mathbf{y})$ a metric. We are now in a position to prove the lemma.

Without loss of generality we may take $\mathbf{y} = (0, 0, \cdots, 0)$ so that $\mathbf{y}' = (1, 0, 1, 0, \cdots, 1, 0)$. Let $\mathbf{x}_i' = (u_{i1}, v_{i1}, u_{i2}, v_{i2}, \cdots, u_{in}, v_{in})$ then

$$d_i^2 = d^2(\mathbf{x}_i, \mathbf{y}) = \sum_{k=1}^{n} \{(1 - u_{ik})^2 + v_{ik}^2\}.$$

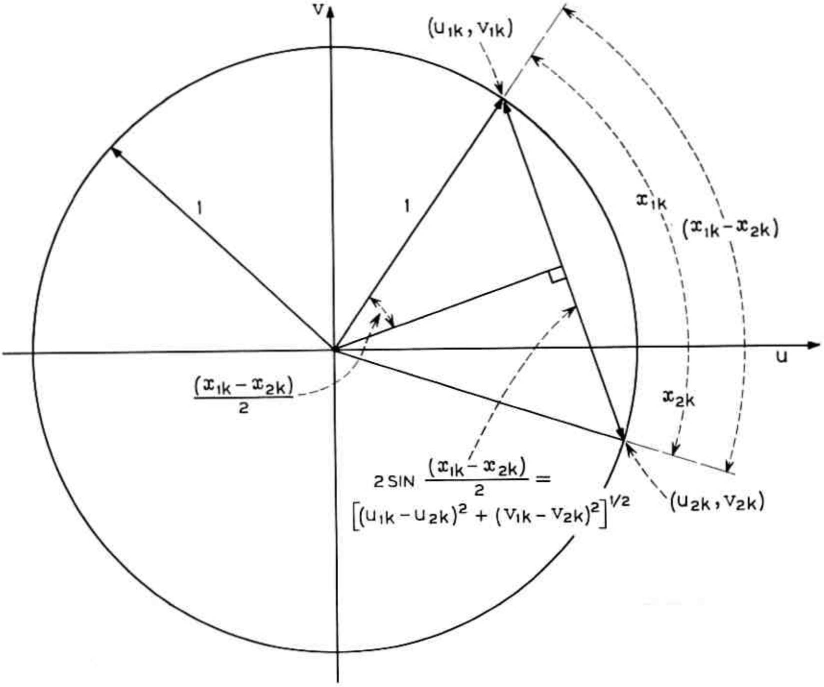

v
(u_{1k}, v_{1k})
1
1
x_{1k}
$(x_{1k} - x_{2k})$
u
$\frac{(x_{1k} - x_{2k})}{2}$
x_{2k}
$2 \sin \frac{(x_{1k} - x_{2k})}{2} = \left[(u_{1k} - u_{2k})^2 + (v_{1k} - v_{2k})^2\right]^{1/2}$
(u_{2k}, v_{2k})

Fig. 6 — Proof of Lemma 2.

Since $d(\mathbf{x}_i, \mathbf{x}_j) \geqq d$,

$$\begin{aligned} \binom{m}{2} d^2 \leqq \sum_{i<j} d^2(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i<j} \sum_{k=1}^{n} \{(u_{ik} - u_{jk})^2 + (v_{ik} - v_{jk})^2\} \\ &= \sum_k \left\{ m \sum_{i=1}^{m} u_{ik}^2 - \left(\sum_{i=1}^{m} u_{ik} \right)^2 \right. \\ &\qquad \left. + m \sum_{i=1}^{m} v_{ik}^2 - \left(\sum_{i=1}^{m} v_{ik} \right)^2 \right\} \\ &= \sum_k \left\{ m \left(\sum_i (1 - u_{ik})^2 \right) \right. \\ &\qquad - \left(m - \sum_i u_{ik} \right)^2 + m \sum_i v_{ik}^2 \\ &\qquad \left. - \left(\sum_i v_{ik} \right)^2 \right\} \\ &= m \sum_i \sum_k \{(1 - u_{ik})^2 + v_{ik}^2\} \end{aligned} \tag{51}$$

$$\begin{aligned} &- \sum_k \left(\sum_i (1 - u_{ik}) \right)^2 \\ &- \sum_k \left(\sum_i v_{ik} \right)^2. \\ \leqq m \sum_i d_i^2 - &\sum_{k=1}^{n} \left(\sum_i (1 - u_{ik}) \right)^2. \end{aligned}$$

From Lemma 1, (51) becomes

$$\binom{m}{2} d^2 \leqq m \sum_i d_i^2 - \frac{1}{n} \left(\sum_{k=1}^{n} \sum_{i=1}^{m} (1 - u_{ik}) \right)^2. \tag{52}$$

Now, since $u_{ik}^2 + v_{ik}^2 = 1$, we have

$$\begin{aligned} d_i^2 &= \sum_k \{(1 - u_{ik})^2 + v_{ik}^2\} = \sum_k \{1 - 2u_{ik} + u_{ik}^2 + v_{ik}^2\} \\ &= 2 \sum_k (1 - u_{ik}). \end{aligned} \tag{53}$$

Substituting (53) into (52) yields

$$\binom{m}{2} d^2 \leqq m \sum_i d_i^2 - \frac{1}{4n} \left(\sum_{i=1}^{n} d_i^2 \right)^2. \tag{54}$$

The lemma follows on multiplying both sides of (54) by $4/n$.

Derivation of the Bound:

If $\mathbf{z} \varepsilon \mathcal{A}_n$ let us define the "sphere" $S(\mathbf{z},\rho)$ as

$$S(\mathbf{z},\rho) = \{\mathbf{x} \varepsilon \mathcal{A}_n : \ d(\mathbf{x},\mathbf{z}) < \rho\}. \tag{55}$$

Since the distance d defined on $\mathcal{A}_n$ is a metric, it follows that if a code $\{\mathbf{x}_i\}_{i=1}^{M}$ has minimum distance (as defined by d), then the spheres $S(\mathbf{x}_i, d/2)$ are disjoint.

Consider the maximum size n-dimensional code with minimum distance d and $M(n,d)$ code words $\{\mathbf{x}_i\}_{i=1}^{M}$. Consider the spheres $S(\mathbf{x}_i, \gamma d)$ about each code word, where

$$\gamma^2 = \frac{1}{\beta}(1 - \sqrt{1 - \beta}) \tag{56a}$$

$$\beta = d^2/2n. \tag{56b}$$

Note that since $\gamma > \frac{1}{2} (0 \leqq \beta \leqq 1)$,* these spheres are not necessarily

* This follows immediately when we write $\gamma^2 = 1/1 + (1 - \beta)^{1/2}$, so that γ increases from $1/\sqrt{2}$ to 1 as β increases from 0 to 1.

disjoint. To each point in the sphere at distance r from the center assign a density $\sigma(r) = \gamma^2 d^2 - r^2$. Then the "mass" of each sphere is

$$\mu = \int_{r<\gamma d} (\gamma^2 d^2 - r^2) dV, \tag{57}$$

where the integration in (57) is performed with respect to the Euclidean measure, assigned to $\mathcal{A}_n$ in the obvious way.

In general, a vector $\mathbf{y} \varepsilon \mathcal{A}_n$ will belong to the spheres about m code words say $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$. We assign to $\mathbf{y}$, a density equal to the sum of the densities contributed by each sphere, i.e.,

$$\sigma_{\mathbf{y}} = \sum_{i=1}^{m} \sigma(d_i) = m\gamma^2 d^2 - \sum_{i=1}^{m} d_i^{\,2}, \tag{58}$$

where $d_i = d(\mathbf{y},\mathbf{x}_i)$. If $\mathbf{y}$ belongs to no sphere $\sigma_{\mathbf{y}} = 0$. Thus, we have

$$\text{mass of } \mathcal{A}_n = \int_{\mathbf{y}\epsilon\mathcal{A}_n} \sigma_{\mathbf{y}} dV = M(n,d)\cdot\mu. \tag{59}$$

We will bound $M(n,d)$ by finding an upper bound on the mass of $\mathcal{A}_n$.

Letting $s = s_{\mathbf{y}} = \sigma_{\mathbf{y}}/n$, (58) becomes

$$\frac{\sum d_i^{\,2}}{n} = \frac{m\gamma^2 d^2}{n} - \frac{\sigma_{\bar{y}}}{n} = 2m\gamma^2\beta - s, \tag{60}$$

where $\beta = d^2/2n$. Substituting (60) into (48) we get

$$(2m\gamma^2\beta - s)^2 - 4m(2m\gamma^2\beta - s) + 4(m)\,(m-1)\beta \leqq 0. \tag{61}$$

Rewriting (61)

$$0 \leqq s^2 \leqq m\{4\beta - 2m\beta(2\gamma^4\beta - 4\gamma^2 + 2) - 4s(1 - \gamma^2\beta)\}. \tag{62}$$

With γ chosen by (56), $2\gamma^4\beta - 4\gamma^2 + 2 = 0$ and $1 - \gamma^2\beta > 0$, so that (62) can only be satisfied if

$$s = \frac{\sigma}{n} \leqq \beta/(1 - \gamma^2\beta) \overset{\Delta}{=} K(\beta). \tag{63}$$

Hence, from (63) and (59) we have

$$M(n,d) = \frac{1}{\mu}\int_{\mathcal{A}_n} \sigma_{\bar{y}} dV \leqq \frac{K(\beta)n}{\mu}\,(\text{Volume of } \mathcal{A}_n). \tag{64}$$

Now from (57)

$$\mu = \int_{r<\gamma d/2} (\gamma^2 d^2 - r^2) dV > \int_{r<\sqrt{\gamma^2 d^2 - 1}} dV = V_n\,(\sqrt{\gamma^2 d^2 - 1}) \tag{65}$$

where $V_n(r)$ is the volume of the sphere in $\mathcal{A}_n$ $S(\mathbf{z},r)$, which is independent of $\mathbf{z}$ (due to the symmetry of $\mathcal{A}_n$). Thus, (64) becomes

$$M(n, \sqrt{2\beta n}) \leqq \frac{nK(\beta)(2\pi)^n}{V_n(\sqrt{\gamma^2 d^2} - 1)}.$$

The asymptotic rate $\hat{R}(\beta)$ satisfies

$$\begin{aligned}\hat{R}(\beta) &= \operatorname*{limit}_{n\to\infty} \frac{1}{n} \ln M(n, \sqrt{2\beta n}) \\ &\leqq \lim_{n\to\infty} \frac{1}{n} \ln \frac{nK(\beta)(2\pi)^n}{V_n(\sqrt{2\gamma^2\beta n} - 1)} \overset{\Delta}{=} \hat{R}_U(\beta).\end{aligned} \tag{66}$$

Applying the result of Appendix C we have $\hat{R}(\beta) \leqq C_0(\gamma^2\beta)$ establishing the upper bound.

4.2 *Lower Bound on $R(\beta)$*

Again let us consider a maximum size n-dimensional code with minimum distance d and $M(n,d)$ code words. About each of the code words $\mathbf{x}_i (i = 1, 2, \cdots, M)$ consider the spheres $S_n(\mathbf{x}_i, d)$. We claim that the union of these spheres $\bigcup_{i=1}^{M} S_n(\mathbf{x}_i, d)$ covers the entire space $\mathcal{A}_n$. This follows from the fact that if $\mathbf{x}_0 \in \mathcal{A}_n$ is in no $S_n(\mathbf{x}_i, d)$, then $d(\mathbf{x}_0, \mathbf{x}_i) \geqq d$, $i = 1, 2, \cdots, M$, so that $\mathbf{x}_0$ may be added to the code destroying the maximality. If $V_n(d)$ is the volume of $S_n(\mathbf{x}_i, d)$ (independent of $\mathbf{x}_i$), then

$$M \cdot V_n(d) \geqq \text{volume of } \mathcal{A}_n = (2\pi)^n. \tag{67}$$

Thus, our lower bound is

$$M(n,d) \geqq \frac{(2\pi)^n}{V_n(d)}. \tag{68}$$

The asymptotic rate $\hat{R}(\beta)$ satisfies

$$\hat{R}(\beta) = \operatorname*{limit}_{n\to\infty} \frac{1}{n} \ln M(n, \sqrt{2\beta n}) \geqq \operatorname*{limit}_{n\to\infty} \frac{1}{n} \ln \frac{(2\pi)^n}{V_n(\sqrt{2\beta n})} \overset{\Delta}{=} \hat{R}_L(\beta). \tag{69}$$

Again applying the result of Appendix C, we have $R(\beta) \geqq C_0(\beta)$ establishing the lower bound.

APPENDIX A

Asymptotic Estimates of the Channel Capacity

The channel capacity C is given by (12) as

$$\frac{C}{W} = -\int_0^\infty \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho + \ln \frac{2A}{e} \tag{70}$$

where

$$\hat{f}(\rho) = 2A\rho e^{-A(1+\rho^2)} I_0(2A\rho). \tag{71}$$

In this appendix we obtain estimates of C for large and small signal-to-noise ratio A.

A.1 *Large A:* We show here that

$$\frac{C}{W} = \tfrac{1}{2} \ln \frac{4\pi A}{e} + \varepsilon_1(A), \tag{72}$$

where $\varepsilon_1(A) \to 0$ as $A \to \infty$. To prove this we will show that for large A nearly all the contribution to the integral in (70) is for ρ in the neighborhood of unity. Part (i) is an estimate of this contribution. Part (ii) shows that the remaining contribution vanishes as $A \to \infty$.

(i) We shall show that if $\delta = A^{-\frac{1}{4}}$,

$$T(A) \triangleq -\int_{1-\delta}^{1+\delta} \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho \to \tfrac{1}{2} \ln \frac{\pi e}{A}, \tag{73}$$

as $A \to \infty$.

Using the asymptotic formula for $\mathrm{I}_0(x)$ for large argument*

$$I_0(x) = \frac{e^x}{\sqrt{2\pi x}} \left[1 + 0\left(\frac{1}{x}\right)\right], \tag{74}$$

we obtain from (71) (for large A)

$$f(\rho) = \sqrt{\frac{A}{\pi}}\, \rho^{\frac{1}{2}} e^{-A(\rho-1)^2} \left[1 + 0\left(\frac{1}{A}\right)\right] \qquad 1 - \delta \leqq \rho \leqq 1 + \delta. \tag{75}$$

Substituting into (73) yields (after a change of variable)

$$T(A) = \left[-1 + 0\left(\frac{1}{A}\right)\right] [B_1 + B_2 + B_3] \tag{76}$$

* Ref. 5, p. 86.

where

$$B_1 = \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}(1+x)^{\frac{1}{2}} \ln \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}\, dx \tag{76a}$$

$$B_2 = \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}(1+x)^{\frac{1}{2}} \ln (1+x)^{\frac{1}{2}}\, dx \tag{76b}$$

$$B_3 = \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}(1+x)^{\frac{1}{2}} 0\left(\frac{1}{A}\right) dx. \tag{76c}$$

Noting that the range of integration is $-\delta \leqq x \leqq \delta$ we can write

$$B_1 = K_1 \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2} \ln \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}\, dx \tag{77a}$$

$$B_2 = K_2 \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}\, dx \leqq K_2 \tag{77b}$$

$$B_3 = K_3 0\left(\frac{1}{A}\right) \int_{-\delta}^{\delta} \sqrt{\frac{A}{\pi}}\, e^{-Ax^2}\, dx \leqq K_3 0\left(\frac{1}{A}\right) \tag{77c}$$

where $(1-\delta)^{\frac{1}{2}} \leqq K_1, K_3 \leqq (1+\delta)^{\frac{1}{2}}, |K_2| \leqq (1+\delta)^{\frac{1}{2}} \ln (1+\delta)^{\frac{1}{2}}$ and $\delta = 1/A^{\frac{1}{4}}$. From (77b) and (77c) we see immediately that $B_2, B_3 \to 0$ as $A \to \infty$ so that we need consider only B_1. From (77a) (letting $y = \sqrt{2A}x$) and setting $\delta = A^{-\frac{1}{4}}$, we have

$$B_1 = \tfrac{1}{2} K_1 \ln \frac{A}{\pi} \int_{-\sqrt{2}A^{\frac{1}{4}}}^{\sqrt{2}A^{\frac{1}{4}}} \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2}\, dy - \frac{K_1}{2} \int_{-\sqrt{2}A^{\frac{1}{4}}}^{\sqrt{2}A^{\frac{1}{4}}} \frac{y^2}{\sqrt{2\pi}}\, e^{-y^2/2}\, dy. \tag{78}$$

Since both integrals in (78) and K_1 tend to unity as $A \to \infty$, we have $B_1 \to \frac{1}{2} \ln (A/\pi e)$ as $A \to \infty$. Applying these results to (76) yields

$$\lim_{A\to\infty} T(A) = \lim_{A\to\infty} \left(-1 + 0\left(\frac{1}{A}\right)\right)\left(\tfrac{1}{2} \ln \frac{A}{\pi e} + B_2 + B_3\right) = \tfrac{1}{2} \ln \frac{\pi e}{A}$$

which is (73).

(*ii*) Here we shall show that with $\delta = A^{-\frac{1}{4}}$ as in (*i*) above,

$$\eta(A) \triangleq \int_{\substack{\rho \leqq 1-\delta \\ \rho \geqq 1+\delta}} \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho \to 0, \quad \text{as} \quad A \to \infty. \tag{79}$$

To do this we write

$$\begin{aligned} \eta(A) = \int_0^{\alpha} \hat{f}(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho + \int_{\alpha}^{1-\delta} f(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d \\ + \int_{1+\delta}^{\infty} f(\rho) \ln \frac{\hat{f}(\rho)}{\rho}\, d\rho = C_1 + C_2 + C_3, \end{aligned} \tag{80}$$

where $\alpha(0 < \alpha < \frac{1}{2})$ is arbitrary. We will show that for arbitrary $\varepsilon > 0$, we can choose A sufficiently large so that $\eta(A) < \varepsilon$. Let us consider each of the integrals C_1, C_2 and C_3 of (80) in turn.

C_1 : For $0 \leqq \rho \leqq \alpha$ we may write

$$\frac{\hat{f}(\rho)}{\rho} \leqq 2Ae^{-A}I_0(2A\alpha), \tag{81}$$

since $I_0(x)$ is an increasing function of x. Making use of the asymptotic formula for $I_0(x)$ (74) we obtain from (81)

$$\left|\frac{\hat{f}(\rho)}{\rho}\right| < \sqrt{\frac{A}{\alpha\pi}}\, e^{-A(1-2\alpha)}\left(1 + 0\left(\frac{1}{A}\right)\right) \to 0, \quad \text{as} \quad A \to \infty,$$

since $\alpha < \frac{1}{2}$. Thus, with A sufficiently large,

$$\left|\frac{\hat{f}(\rho)}{\rho} \log \frac{\hat{f}(\rho)}{\rho}\right| \leqq \frac{2\varepsilon}{3\alpha^2}, \qquad 0 \leqq \rho \leqq \alpha$$

and

$$|\, C_1 \,| \leqq \int_0^\alpha \frac{2\varepsilon}{3\alpha^2}\, \rho \, d\rho = \frac{\varepsilon}{3}. \tag{82}$$

C_2 : Again using the asymptotic formula for $I_0(x)$ (74) we may write, for $\alpha \leqq \rho \leqq 1 - \delta$,

$$\begin{aligned}\frac{\hat{f}(\rho)}{\rho} &= \sqrt{\frac{A}{\pi\alpha}}\, e^{-A(1-\rho)^2}\left(1 + 0\left(\frac{1}{A}\right)\right) \leqq \sqrt{\frac{A}{\pi\alpha}}\, e^{-A\delta^2}\left(1 + 0\left(\frac{1}{A}\right)\right)\\ &= \sqrt{\frac{A}{\pi\alpha}}\, e^{-A^{\frac{1}{3}}}\left(1 + 0\left(\frac{1}{A}\right)\right) \to 0 \quad \text{as} \quad A \to \infty.\end{aligned} \tag{83}$$

Thus, with A sufficiently large

$$\left|\frac{\hat{f}(\rho)}{\rho} \ln \frac{\hat{f}(\rho)}{\rho}\right| \leqq \tfrac{2}{3}\varepsilon,$$

from which

$$|\, C_2 \,| \leqq \tfrac{2}{3}\varepsilon \int_\alpha^{1-\delta} \rho \, d\rho \leqq \frac{\varepsilon}{3}. \tag{84}$$

C_3 : As above, we may write for $\rho \geqq 1 + \delta$,

$$\frac{\hat{f}(\rho)}{\rho} = \sqrt{\frac{A}{\pi\rho}}\, e^{-A(\rho-1)^2}\left(1 + 0\left(\frac{1}{A}\right)\right). \tag{85}$$

Substituting (85) into the defining integral for C_3 (80), and making change of variable $y = (\rho - 1)$, we obtain

$$\begin{aligned} C_3 = & \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,(1+y)^{\frac{1}{2}}\,e^{-Ay^2}\ln\sqrt{\frac{A}{\pi}}\,dy \\ & + \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,(1+y)^{\frac{1}{2}}\,e^{-Ay^2}\ln\,(1+y)^{-\frac{1}{2}}\,dy \\ & + \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,(1+y)^{\frac{1}{2}}\,e^{-Ay^2}\ln e^{-Ay^2}\,dy \\ & + \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,(1+y)^{\frac{1}{2}}\,e^{-Ay^2}\ln\left(1+0\left(\frac{1}{A}\right)\right)dy. \end{aligned} \tag{86}$$

Since for $y \geqq \delta$, $(1+y)^{\frac{1}{2}} \leqq 2e^{y^2}$ and $|\,(1+y)^{\frac{1}{2}}\ln\,(1+y)^{-\frac{1}{2}}\,| \leqq e^{y^2}$, we have from (86)

$$\begin{aligned} |C_3| \leqq & \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,e^{-(A-1)y^2}\left[\tfrac{1}{2}\ln\frac{A}{\pi}+2+0\left(\frac{1}{A}\right)\right]dy \\ & \qquad + \int_\delta^\infty \sqrt{\frac{A}{\pi}}\,e^{-(A-1)y^2}(Ay^2)\,dy. \end{aligned} \tag{87}$$

Using the well known asymptotic formula for the cumulative error function, and the fact that $\delta = A^{-\frac{1}{4}}$, it is readily shown that for large A,

$$\int_\delta^\infty \sqrt{\frac{A}{\pi}}\,e^{-(A-1)y^2}\,dy \approx \frac{1}{\sqrt{2\pi}A^{\frac{1}{4}}}\,e^{-A^{\frac{1}{2}}}, \tag{88a}$$

and

$$\int_\delta^\infty \sqrt{\frac{A}{\pi}}\,e^{-(A-1)y^2}\,(Ay^2)\,dy \approx \frac{A^{\frac{1}{4}}}{\sqrt{2\pi}}\,e^{-A^{\frac{1}{4}}}. \tag{88b}$$

Equations (88a and b) tell us that with A sufficiently large $|\,C_3\,| \leqq \varepsilon/3$.

Taking the above results together yields

$$|\,\eta(A)\,| \leqq |\,C_1\,| + |\,C_2\,| + |\,C_3\,| \leqq \varepsilon,$$

with A sufficiently large.

(*iii*) The final step is to substitute (73) and (79) into (70) and obtain

$$\frac{C}{W} \to \tfrac{1}{2}\ln\frac{\pi e}{A} + \ln\frac{2A}{e} = \tfrac{1}{2}\ln\frac{4\pi}{e}A, \quad \text{as} \quad A \to \infty,$$

which is what is to be proved, (72).

A.2 *Small A:* We show here that

$$\frac{C}{W} = A(1 + 0(A)), \tag{89}$$

as $A \rightarrow 0$.

Substituting (71) into (70) yields, after a bit of straightforward manipulation,

$$\begin{aligned}\frac{C}{W} = A - 1 + 2A^2e^{-A} \int_0^\infty \rho^3 e^{-A\rho^2}\, I_0(2\rho A)d\rho \\ - 2Ae^{-A} \int_0^\infty \rho e^{-A\rho^2} I_0(2\rho A) \ln I_0(2\rho A)d\rho.\end{aligned} \tag{90}$$

If we change the variable of integration to $x = 2\rho A$, we obtain from (90)

$$\begin{aligned}\frac{C}{W} = A - + \frac{e^{-A}}{8A^2} \int_0^\infty x^3 I_0(x)e^{-x^2/4A}dx \\ - \frac{e^{-A}}{2A} \int_0^\infty xI_0(x)e^{-x^2/4A} \ln I_0(x)dx.\end{aligned} \tag{91}$$

Now the first integral of (90) is known* and is

$$\int_0^\infty x^3 I_0(x)e^{-x^2/4A}\, dx = 8A^2(1 + A)e^A, \tag{92}$$

so that

$$\frac{C}{W} = 2A - \frac{e^A}{2A} \int_0^\infty xI_0(x)e^{-x^2/4A} \ln I_0(x)dx = 2A - \frac{e^{-A}}{2A} D. \tag{93}$$

We can estimate the integral D for small A, by noting that most of the contribution is for small x. Making use of the asymptotic formula for $I_0(x)$, for small x

$$I_0(x) = 1 + \frac{x^2}{4} + 0(x^4), \tag{94}$$

we have

$$D = \int_0^\infty \left[\frac{x^3}{4} + 0(x^5)\right] e^{-x^2/4A}\, dx. \tag{95}$$

* Ref. 6, p. 198, (4a).

Since

$$\int_0^\infty \frac{x^3}{4} e^{-x^2/4A} \, dx = 2A^2,$$

and

$$\int_0^\infty x^5 e^{-x^2/4A} \, dx = 64A^3\Gamma(3) = 0(A^3),$$

we have

$$D = 2A^2(1 + 0(A)). \tag{96}$$

From (96) and (93) we get

$$\frac{C}{W} = 2A - Ae^{-A}(1 + 0(A)) = A(1 + 0(A))$$

which is what was to be proved (89).

APPENDIX B

The Function λ(ξ)

In this appendix, we show that for ξ satisfying

$$0 < \xi \leqq 1, \tag{97}$$

there exists a unique λ(ξ) which satisfies

$$\xi = 1 - \frac{I_1(\lambda(\xi))}{I_0(\lambda(\xi))}. \tag{98}$$

If we define the function ξ (λ) by

$$\xi(\lambda) = 1 - \frac{I_1(\lambda)}{I_0(\lambda)}, \qquad 0 \leqq \lambda < \infty, \tag{99}$$

it will suffice to show that

(*i*) $\xi(\lambda)$ is strictly monotone decreasing,
(*ii*) $\xi(0) = 1$,
(*iii*) $\lim_{\lambda\to\infty} \xi(\lambda) = 0$.

If (*i*), (*ii*), and (*iii*) are true, ξ(λ) is a one-to-one mapping of the half line [0, ∞) onto the interval (0,1].

(*i*) Making use of the fact that $I_0'(\lambda) = I_1(\lambda)$ we can write

$$\frac{d\xi(\lambda)}{d\lambda} = \frac{-I_0(\lambda)I_1'(\lambda) + I_1^2(\lambda)}{(I_0(\lambda))^2}. \tag{100}$$

Since*

$$I_0(\lambda) = \frac{1}{\pi}\int_0^\pi e^{\lambda\cos\varphi}\,d\varphi,$$

we have

$$I_1(\lambda) = I_0'(\lambda) = \frac{1}{\pi}\int_0^\pi \cos\varphi e^{\lambda\cos\varphi}\,d\varphi,$$

and

$$I_1'(\lambda) = \frac{1}{\pi}\int_0^\pi \cos^2\varphi e^{\lambda\cos\varphi}\,d\varphi.$$

Thus (100) becomes

$$\frac{d\xi(\lambda)}{d\lambda} = \frac{-\frac{1}{\pi^2}\int_0^\pi e^{\lambda\cos\varphi}\,d\varphi\int_0^\pi \cos^2\varphi e^{\lambda\cos\varphi}\,d\varphi + \frac{1}{\pi^2}\left(\int_0^\pi \cos\varphi e^{\lambda\cos\varphi}\,d\varphi\right)^2}{[I_0(\lambda)]^2}. \tag{101}$$

By the Schwarz inequality

$$\left(\int_0^\pi \cos\varphi e^{\lambda\cos\varphi}\,d\varphi\right)^2 < \left(\int_0^\pi \cos^2\varphi e^{\lambda\cos\varphi}\,d\varphi\right)\left(\int_0^\pi e^{\lambda\cos\varphi}\,d\varphi\right),$$

(the strict inequality holding). Hence $\frac{d\xi(\lambda)}{d\lambda} < 0$ and (*i*) follows.

$$(ii)\ \xi(0) = 1 - \frac{I_1(0)}{I_0(0)} = 1 - \frac{0}{1} = 1.$$

(*iii*) We make use of the asymptotic formula for $I_0(x)$ and $I_1(x)$ for large x†

$$\begin{aligned} I_0(x) &= \frac{e^x}{\sqrt{2\pi x}}\left[1 + \frac{1}{8x} + 0\left(\frac{1}{x^2}\right)\right] \\ I_1(x) &= \frac{e^x}{\sqrt{2\pi x}}\left[1 - \frac{3}{8x} + 0\left(\frac{1}{x^2}\right)\right]. \end{aligned} \tag{102}$$

Substitution of (102) into (99) yields (*iii*) immediately.

* Ref. 4, p. 76.
† Ref. 4, p. 86.

Let us remark here that since $I_0(x)$ and $I_1(x)$ are even functions of x, if $\lambda(\xi) = a \geqq 0$ is the unique nonnegative solution to (98), then $\lambda(\xi) = -a$ is the unique nonpositive solution to (98).

APPENDIX C

Completion of Asymptotic Estimates of $R(\beta)$

We have defined $V_n(r)$ as the volume of the sphere $S_n(\mathbf{z},r) = \{x \varepsilon \mathcal{A}_n : \ d(\mathbf{x},\mathbf{z}) < r\}$. Due to the symmetry of $\mathcal{A}_n$, $V_n(r)$ is independent of $\mathbf{z}$. Thus, we shall take $V_n(r)$ as the volume of

$$S(\bar{0},\mathbf{r}) = \left\{\mathbf{x} = (x_1, x_2, \cdots, x_n)\ \varepsilon\ \mathcal{A}_n\text{:}\quad d^2(\bar{0},\mathbf{x}) = \sum_{k=1}^{n}\left(2 \sin \frac{x_k}{2}\right)^2 < r^2\right\}.$$

In this appendix, we evaluate

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{(2\pi)^n}{V_n(\sqrt{an})} \triangleq E_a . \tag{103}$$

We shall find E_a by solving an equivalent probability problem: Let $X_1, X_2, \cdots$ be a sequence of independent random variables uniformly distributed on the interval $[-\pi,\pi]$. Let

$$Y_n = \sum_{k=1}^{n}\left(2 \sin \frac{X_k}{2}\right)^2 .$$

It is clear that

$$\Pr[Y_n < r^2] = \frac{V_n(r)}{(2\pi)^n}, \tag{104}$$

hence,

$$-\lim_{n\to\infty}(1/n) \ln \Pr[Y_n < an] = E_a . \tag{105}$$

We now make use of

Chernoff's Theorem:[7] *Let $Z_1, Z_2, \cdots$ be a sequence of independent identically distributed random variables with moment generating function $E[e^{Z_k t}] = M(t)$. Let*

$$P_n = \Pr\left[\sum_{k=1}^{n} Z_i \leqq an\right],$$

where $a \leqq E(Z_k)$. Then

$$\lim_{n\to\infty} \frac{1}{n} \ln P_n = \ln m,$$

where $m = \min_{t \leqq 0} e^{-at} M(t)$.

If we set

$$Z_k = \left[2 \sin \frac{X_k}{2} \right],$$

where X_k is the above random variable, then

$$Y_n = \sum_{k=1}^{n} Z_k .$$

Thus, from (105) and Chernoff's Theorem, $E_a = -\ln m$.

The moment generating function of Z_k is

$$\begin{aligned} M(t) = E[e^{Z_k t}] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp \left\{ \left(2 \sin \frac{x}{2} \right)^2 t \right\} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{(2t - 2t \cos x)} \, dx = e^{2t} I_0(2t). \end{aligned} \tag{106}$$

Hence,

$$m = \min_{t \leqq 0} e^{(2-a)t} I_0(2t). \tag{107}$$

To find the minimum, set the derivative of (107) equal to zero:

$$0 = e^{(2-a)t}[(2 - a)I_0(2t) + 2I_1(2t)],$$

so that the t which minimizes (107) satisfies

$$\frac{a}{2} = 1 - \frac{I_1(2t)}{I_0(2t)} . \tag{108}$$

The solution to (108), for $t \leqq 0$, is $2t = -\lambda(a/2)$, where $\lambda(\xi)$ is defined by (12). (See the remark at the conclusion of Appendix B.) Hence, from (107)

$$m = \exp \left[-\left(1 - \frac{a}{2} \right) \lambda \left(\frac{a}{2} \right) \right] I_0 \left(\lambda \left(\frac{a}{2} \right) \right), \tag{109}$$

so that

$$\begin{aligned} E_a = -\ln m &= -\ln I_0 \left(\lambda \left(\frac{a}{2} \right) \right) + \left(1 - \frac{a}{2} \right) \lambda \left(\frac{a}{2} \right) \\ &= C_0 \left(\frac{a}{2} \right) \end{aligned} \tag{110}$$

where $C_0(\xi)$ is defined by (16).

Applying (110) to (66) yields $\hat{R}_U(\beta) = C_0(\gamma^2\beta)$, and applying (110) to (69) yields $\hat{R}_L(\beta) = C_0(\beta)$.

APPENDIX D

Exponential Error Bounds

It is known that for any time-discrete (amplitude continuous) memoryless channel the smallest attainable error probability $P_e^*(n,\hat{R})$ for an n-dimensional code with $e^{n\hat{R}}$ code words may be written

$$P_e^*(n,\hat{R}) = \exp\left[-nE(\hat{R}) + o(n)\right], \tag{111}$$

where $E(\hat{R}) > 0$ when $\hat{R} < \hat{C}$ (the channel capacity in nats per symbol). Although $E(\hat{R})$ is not always known exactly it can be estimated by upper and lower bounds. The best known lower bound on $E(\hat{R})$ is given in Gallager (Ref. 8, Theorem 10) and the best known upper bound on $E(\hat{R})$ by Shannon, Gallager, and Berlekamp.[9]

Let $P(y \mid x)$ be the channel transition probability density. We assume that any n-sequence of input symbols is an allowable channel input — i.e., no "input constraint". The bounds of Refs. 8 and 9 can then be stated as follows:

For any $\rho \geqq 0$ and input probability density $f(x)$, let us define

$$E(\rho,f) = E_0(\rho,f(x)) - \rho R_o(\rho,f(x)), \tag{112}$$

where

$$E_0(\rho,f(x)) = -\ln \int_y dy \left[\int_x dx\, f(x) P(y \mid x)^{1/1+\rho} \right]^{1+\rho} \tag{112a}$$

and

$$R_o(\rho,f(x)) = \frac{\partial}{\partial\rho} E_0(\rho,f(x)). \tag{112b}$$

With $\rho \geqq 0$ specified let $f_\rho(x)$ be that input density which maximizes $E(\rho.f(x))$. It is shown in Ref. 8 that with ρ fixed $f_\rho(x)$ is the unique density which satisfies

$$\int_y dy\, P(y \mid x)^{1/1+\rho} \alpha_\rho^{\rho}(y) \geqq \int_y \alpha_\rho(y)^{1+\rho} dy, \qquad \text{all } x, \tag{113}$$

with equality if $f\rho(x) \neq 0$ (all x) where

$$\alpha_\rho(y) = \int_x f_\rho(x) P(y \mid x)^{1/1+\rho} dx. \tag{113a}$$

It can be shown that with $\rho = 0$, $\mathrm{f}_0(x)$ is that input density which achieves capacity $\hat{C}$, and $R_o(0,f_0) = \hat{C}$. In most channels of interest $R_o(\rho,f_\rho)$ decreases from $\hat{C}$ to 0 as ρ increases from 0 to ∞.

We define the rate $\hat{R}$ parametrically in terms of ρ by

$$\hat{R} = \hat{R}(\rho) = R_o(\rho,f_\rho(x)). \tag{114}$$

Then for $0 \leqq \rho \leqq 1$, which corresponds to $\hat{R}_o(1,f_1(x)) \leqq \hat{R} \leqq \hat{C}$, the exponent is known exactly:

$$E(\hat{R}) = E(\rho,f_\rho) = E_0(\rho,f_\rho) - \rho\hat{R} \tag{115}$$

where E, E_0, and f_ρ are defined by (112). For $\rho \geqq 1$, which corresponds to $\hat{R} \leqq \hat{R}_o(1,f_1)$, the ("sphere-packing"), upper bound on $E(\hat{R})$ is

$$E(\hat{R}) \leqq E_0(\rho,f_\rho), \tag{116}$$

and the ("random-coding") lower bound is for $0 \leqq \hat{R} \leqq \hat{R}_o(1,f_1)$

$$\mathrm{E}(\hat{R}) \geqq E_0(1,f_1) - \hat{R}. \tag{117}$$

This estimate of $E(\hat{R})$ may be improved for low rates $\hat{R}$. It is shown in Ref. 9 that if $E^*(\hat{R})$ is an upper bound on $E(\hat{R})$ which is sharper than the sphere-packing bound (116) for low rates $\hat{R}$ (such a bound can always be found), and if $E^*(\hat{R})$ and the sphere-packing bound are plotted versus $\hat{R}$, then their common tangent is also an upper bound on $E(\hat{R})$.

The lower bound may be sharpened for low rates $\hat{R}$ as follows. For $\rho \geqq 1$, and input density $g(x)$, define

$$E_x(\rho,g) = E_{0x}(\rho,g) - \rho R_{0x}(\rho,g), \tag{118}$$

where

$$E_{0x}(\rho,g)$$

$$= -\rho \ln \int_x g(x)dx \int_{x'} g(x')dx' \left[\int_y P(y \mid x)^{\frac{1}{2}} P(y \mid x')^{\frac{1}{2}} dy\right]^{1/\rho} \tag{118a}$$

and

$$R_{0x}(\rho,g) = \frac{\partial}{\partial\rho} E_{0x}(\rho,g). \tag{118b}$$

Then for any fixed $g(x)$ and with $\hat{R}$ again given parametrically in terms of ρ by

$$\hat{R} = R_{0x}(\rho,g), \tag{119}$$

the ("expurgated") lower bound is

$$E(R) \geqq E_x(\rho,g). \tag{120}$$

We shall now apply these results to our channel using the time-discrete model defined before and after (26). Here the input is a number $X \,\varepsilon\, [-\pi,\pi]$, and the output is a pair of real numbers (Y_1, Y_2). If $X = x$ is the input, then the conditional transition probability density is the two-dimensional

$$P(y_1, y_2 \mid x) = \frac{1}{2\pi N} \exp \{-[(y_1 - \sqrt{2S} \cos x)^2 + (y_2 - \sqrt{2S} \sin x)^2]/2N\}$$

or in polar coordinates

$$\begin{aligned} P(r,\varphi \mid x) &= rP(r \cos \varphi, r \sin \varphi \mid x) \\ &= \frac{r}{2\pi N} e^{-S/N} e^{-r^2/2N} \exp\left(-r \frac{\sqrt{2S'}}{N} \cos (\varphi - x)\right). \end{aligned} \tag{121}$$

It may be verified by substitution into (113), that the input density $f(x) = 1/2\pi$ maximizes $E(\rho,f(x))$ for all $\rho \geqq 0$. Further a direct substitution of (121) into (112a) yields after a straightforward computation

$$E_0(\rho,f\rho) = -\ln 2Ae^{-A} \int_0^\infty ve^{-Av^2} \left[I_0 \left(\frac{2Av}{1+\rho} \right) \right]^{1+\rho} dv, \tag{122}$$

where $A = S/N$, the signal-to-noise ratio. The rate, $\hat{R}$, can be gotten by differentiating (122) with respect to ρ. This yields

$$\begin{aligned} \hat{R}(\rho) &= \frac{\partial}{\partial \rho} E_0(\rho,f_\rho) \\ &= \left[-\int_0^\infty ve^{-Av^2} I_0 \left(\frac{2vA}{1+\rho} \right)^{1+\rho} \ln I_0 \left(\frac{2vA}{1+\rho} \right) dv \right. \\ &\quad \left. + \frac{2A}{(1+\rho)} \int_0^\infty v^2 e^{-Av^2} I_0 \left(\frac{2Av}{1+\rho} \right)^\rho I_1 \left(\frac{2Av}{1+\rho} \right) dv \right] \Big/ \\ &\quad \int_0^\infty ve^{-Av^2} I_0 \left(\frac{2Av}{1+\rho} \right)^{1+\rho} dv. \end{aligned} \tag{123}$$

After some manipulation one can show that $\hat{R}(\rho) \mid_{\rho=0} = \hat{C}$, the channel capacity as given by (12). The estimate of the exponent $E(\hat{R})$ of (115), (116), and (117) is plotted versus $\hat{R}$ in Fig. 7 for signal-to-noise ratio

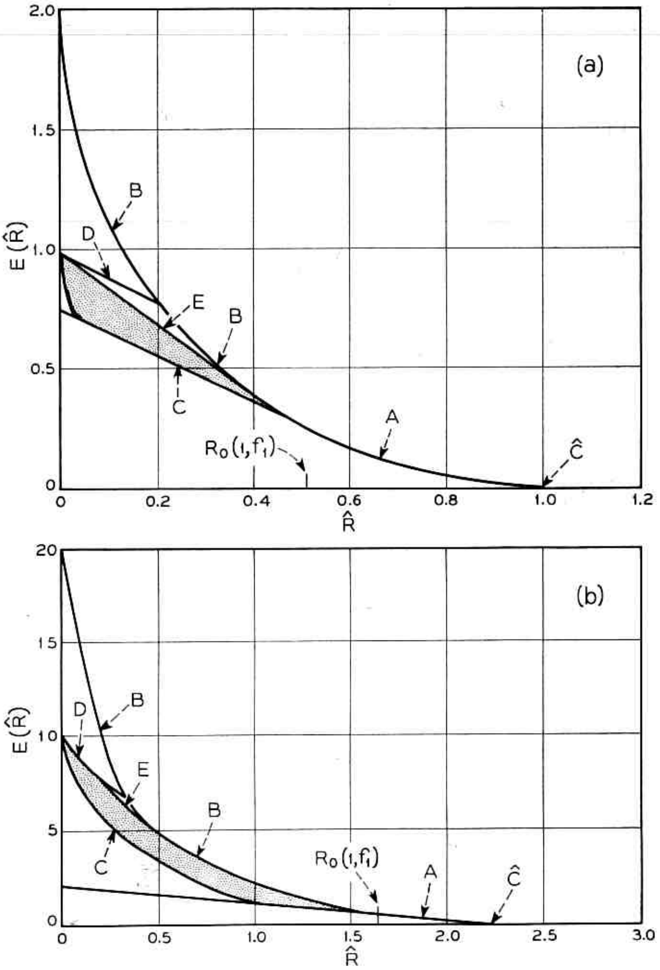


Fig. 7 — Upper and lower bounds on the error exponent $E(\hat{R})$ vs. $\hat{R}$ for signal-to-noise ratios of (a) $A = 2$, (b) $A = 20$. (Curve A is the exponent $E(\hat{R})$ in the range where it is known exactly (115). Curve B is the "sphere-packing" upper bound on $E(\hat{R})$ (116). Curve $\hat{C}$ is the "random coding" lower bound on $E(\hat{R})$ (117). Curve D is $E^*(\hat{R})$, the low rate upper bound (133). Curve E is the common tangent to $E^*(\hat{R})$ and the sphere-packing bound. $E(\hat{R})$ lies in the shaded region.

$A = 2,20$. The sphere-packing upper bound and the random-coding lower bound diverge for small rates $\hat{R}$. We shall improve this situation by computing the low-rate expurgated lower bound on $E(\hat{R})$ (120). If we again choose the input density to be $g(x) = 1/2\pi$, $-\pi \leqq x \leqq \pi$ we have from (121)

$$\int_y P(y\,|\,x)^{\frac{1}{2}}P(y\,|\,x')^{\frac{1}{2}}dy = \int_0^\infty dr\,\frac{r}{2\pi N}\exp\left(-S/N_e - r^2/2N\right)$$

$$\cdot\int_{-\pi}^{\pi}\exp\left(-\frac{r}{N}\sqrt{\frac{S}{2}}\,[\cos(\varphi - x) + \cos(\varphi - x')]\right)d\varphi$$

$$= \int_0^\infty dr\,\frac{r}{2\pi N}\exp\left(-S/N - r^2/2N\right)$$

$$\cdot\int_{-\pi}^{\pi}\exp\left(-\frac{\sqrt{S}}{N}\,Br\cos(\varphi - a)\right)d\varphi$$

$$= \frac{\exp(-S/N)}{N}\int_0^\infty r\exp(-r^2/2N)I_0\left(\frac{\sqrt{S}}{N}\,Br\right)dr,$$

where

$$B = \sqrt{1 + \cos(x - x')} \quad \text{and} \quad a = \tan^{-1}\left[\frac{\sin x + \sin x'}{\cos x + \cos x'}\right].$$

This integral is tabulated [Ref. 6, p. 198, #5] so that we have

$$\int_y P(y\,|\,x)^{\frac{1}{2}}P(y\,|\,x')^{\frac{1}{2}}dy = \exp\left(-(S/2N)[1 - \cos(x - x')]\right). \qquad (124)$$

Substituting (124) into (118a) yields

$$\begin{aligned} E_{0x}(\rho,g) &= -\rho\ln\int_{-\pi}^{\pi}\frac{dx}{2\pi}\int_{-\pi}^{\pi}\frac{dx'}{2\pi} \\ &\qquad\cdot\exp\left[-S/2N\rho - (S/2N\rho)\cos(x - x')\right] \\ &= -\rho\ln\int_{-\pi}^{\pi}\frac{dx}{2\pi}\exp(-S/2N\rho)I_0\left(\frac{S}{2N\rho}\right) \\ &= -\rho\ln\left[\exp(-A/2\rho)I_0(A/2\rho)\right], \qquad (\rho \geqq 1) \end{aligned} \qquad (125)$$

where $A = S/N$. The rate $\hat{R}$ is given parametrically in terms of ρ by

$$\hat{R} = \hat{R}(\rho) = \frac{\partial E_{0x}}{\partial\rho}(\rho,g) = \frac{A}{2\rho}\,\frac{I_1(A/2\rho)}{I_0(A/2\rho)} - \ln I_0(A/2\rho) \qquad (126)$$

$$(\rho \geqq 1).$$

Let us note that as $\rho \to \infty$ $\hat{R}(\rho) \to 0$. The expurgated bound is given by (120), (125), and (126). It is easy to show that as $\rho \to \infty$, $(\hat{R} \to 0)$ the lower bound $E_x(\rho,g) \to A/2$.

We shall now obtain a sharper upper bound for low rates $E^*(\hat{R})$ which will in fact have $E^*(0) = A/2$, establishing that $E(0) = A/2$.

Let us denote by $\rho_n(M)$, the smallest maximum (normalized) correlation obtainable for an n-dimensional polyphase code with M code words. Paralleling arguments of Shannon (Ref. 10, pp. 647–648) it is not hard to show that the error probability for a code with $M = \exp\,(n\hat{R})$ code words satisfies

$P_e \geqq \frac{1}{2}\mathrm{Pr}$ [error in a code with two code words with energy ST and correlation $\rho_n(M/2)$ in white Gaussian noise with spectral density N_o].

The right member of this inequality is known [Ref. 11, (38)], and is equal to

$$\tfrac{1}{2}\Phi\left(-\sqrt{\frac{ST}{N_0}\left(1-\rho_n\left(\frac{M}{2}\right)\right)}\right), \tag{127}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}du$$

is the cumulative error function.

We now bound $E(\hat{R})$ by finding a bound on $\rho_n(M/2)$. Since for large M, a code with $M/2$ code words has about the same rate as one with M code words, it will suffice to bound $\rho_n(M)$. With $M = \exp\,(n\hat{R})$ and n large, we have from (17) (since $\beta = (1-\rho)$)

$$\hat{R} \leqq C_0(1-\sqrt{\rho_n(M)}). \tag{128}$$

If we define $\hat{R}$ parametrically by

$$\hat{R} = \hat{R}(\sigma) = C_0(\sigma), \qquad 0 \leqq \sigma \leqq 1, \tag{129}$$

we have from (129)

$$\sigma \geqq 1 - \sqrt{\rho_n(M)} \tag{130}$$

or

$$[1-\rho_n(M)] \leqq \sigma(2-\sigma). \tag{131}$$

Substituting (131) into (127) yields

$$P_e \leqq \tfrac{1}{2}\Phi\left(-\sqrt{\frac{ST}{N_o}\,\sigma(2-\sigma)}\right). \tag{132}$$

Making use of the well-known asymptotic formula for the cumulative error function $\Phi(-x) \approx (1/\sqrt{2\pi}x)\exp\,(-x^2/2)$ (large x), we obtain from (132) for large T (and therefore large $n = WT$), the upper bound on the error exponent

$$E(\hat{R}) = -\lim_{n\to\infty} \frac{1}{n} \ln P_e \leqq \frac{A}{2}\sigma(2-\sigma) \triangleq E^*(R) \tag{133}$$

where $A = S/N_oW = S/N$. When $R = 0$, $\sigma = 1$, so that $E(0) = A/2$. The expurgated bound and the bound of (133) are plotted in Fig. 7. The upper bound is, of course, sharpened by drawing the common tangent of $E^*(\hat{R})$ and the sphere-packing bound.

APPENDIX E

Asymptotic Estimates of $C_0(\xi)$

In this appendix we obtain estimates of $C_0(\xi)$ as $\xi \to 0$ and $\xi \to 1$.

E.1 Small ξ: We show here that

$$C_0(\xi) = \tfrac{1}{2} \ln \frac{\pi}{e\xi} + \varepsilon_1(\xi) \tag{134}$$

where $\varepsilon_1(\xi) \to 0$ as $\xi \to \infty$.

From proposition (*iii*) in Appendix C, we know that as $\xi \to 0$, $\lambda(\xi) \to \infty$. Again making use of the asymptotic formula for $I_0(x)$ and $I_1(x)$ for large x (102), we obtain by substitution into (15),

$$\xi = \frac{1}{2\lambda}\left[1 + 0\left(\frac{1}{\lambda}\right)\right]. \tag{135}$$

or

$$\lambda = \frac{1}{2\xi} + 0(1) = \frac{1}{2\xi} + k + \varepsilon(\xi), \tag{136}$$

where $\varepsilon(\xi) \to 0$ as $\xi \to 0$ and k is a constant. Substitution of (136) into (16), and another application of the asymptotic formula for $I_0(x)$ yields (134).

E.2 Large ξ: We show here that

$$C_0(\xi) = (1-\xi)^2\{1 + 0[(1-\xi)^2]\} \tag{137}$$

as $\xi \to 1$. As above our first task is to estimate $\lambda(\xi)$ when ξ is near unity or $\lambda(\xi)$ is near zero. We need the asymptotic formulas for $I_0(x)$ and $I_1(x)$ for x near zero (Ref. 4, p. 77):

$$\begin{aligned} I_0(x) &= 1 + \frac{x^2}{4} + 0(x^4), \\ I_1(x) &= \frac{x}{2} + \frac{x^3}{16} + 0(x^5). \end{aligned} \tag{138}$$

Substituting (138) into (15) yields

$$\xi = 1 - \frac{I_1(\lambda)}{I_0(\lambda)} = 1 - \frac{\lambda}{2} + 0(\lambda^3). \tag{139}$$

Setting $\hat{\xi} = 1 - \xi$ we have,

$$\hat{\xi} = \frac{\lambda}{2} + 0(\lambda^3). \tag{140}$$

We show that

$$\lambda = 2\hat{\xi} + 0(\hat{\xi}^3) = 2(1 - \xi) + 0((1 - \xi)^3). \tag{141}$$

Equation (141) follows on setting $x = \lambda - 2\xi$ and observing [from (140)] that

$$\frac{x}{\hat{\xi}^3} = \frac{0(\lambda^3)}{\left[\frac{\lambda}{2} + 0(\lambda)^3\right]^3} \to k,$$

as $\lambda \to 0$ or $\hat{\xi} \to 0$ $(\xi \to 1)$. Substitution of (141) into (16) and another application of the asymptotic formula for $I_0(x)$, yields (137).

APPENDIX F

Comparison of Modulation Schemes

In this appendix, we shall describe an amplitude modulation scheme and compare its performance with that of the phase modulation scheme studied in this paper.

Referring to (21a) we see that our phase modulated signal may be written (during the kth subinterval)

$$s_i(t) = x_{ik}^{(1)} \sin \alpha 2\pi Wt + x_{ki}^{(2)} \cos \alpha 2\pi Wt, \tag{142}$$

where from (21b)

$$[x_{ik}^{(1)}]^2 + [x_{ij}^{(2)}]^2 = 2S, \qquad k = 1, 2, \cdots, n. \tag{143}$$

Consider an amplitude modulation (AM) scheme in which the signals $s_i(t)$ are given by (142) but with (143) replaced by the "mean square" constraint

$$\sum_{k=1}^{n} \{[x_{ik}^{(1)}]^2 + [x_{ik}^{(2)}]^2\} \leqq 2Sn = 2WST. \tag{144}$$

The resulting signals $s_i(t)$ are then amplitude modulated signals with carrier frequency $\alpha 2\pi W$ radians per second and average power

$$\frac{1}{T}\int_0^T s_i^2(t)dt = \frac{1}{T}\sum_{k=1}^{n}\int_{(k-1)T/n}^{kT/n} s_i^2(t)dt \leqq S. \tag{145}$$

Thus, in this case the signals are constrained to have average power not exceeding S. It is clear that, as for the phase modulation, the signals of the αth and βth users of the channel are orthogonal so that we may again take the bandwidth (i.e., difference in carrier frequencies of adjacent users) to be W cps. Further, it follows from the analysis in Section III that this channel is mathematically equivalent to the time-discrete channel Gaussian channel considered by Shannon.[10,12] This channel accepts real numbers at a rate of $2W$ per second and adds to each number an independent Gaussian variate with mean zero and variance N_0W. Messages are encoded in blocks (vectors) of $2WT$ real numbers (which take T seconds to transmit), each $2WT$-vector having the sum of the squares of the coordinates not exceeding $2WST$. Shannon has found the capacity of this channel to be (in nats per second)

$$W \ln\left(1 + \frac{S}{N_0W}\right) = W \ln (1 + A), \tag{146}$$

where $A = S/N_0W$, the signal to noise ratio. Equation (146) is plotted in Fig. 2 so that it may be compared to the capacity of the polyphase system. Note that for small A, ln $(1 + A) \approx A$ so that from (14) the

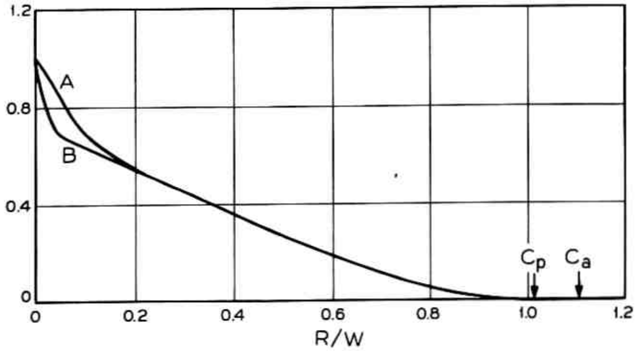


Fig. 8 — Lower bound on the exponents $E_a(R)$ (curve A) and $E_p(R)$ curve B). C_p and C_a are the capacities of the polyphase and AM systems, respectively. The signal-to-noise ratio A=2.

capacities of this AM scheme and the polyphase scheme are nearly the same.

Further, letting $P_{ep}{}^*(T,R)$ and $P_{ea}{}^*(T,R)$ be the smallest attainable error probability for a code with parameter T and rate R nats per sec. for the polyphase and AM systems, respectively, we can write

$$P_{ep}{}^* = \exp\left[-TE_p(R) + 0(T)\right]$$

$$P_{ea}{}^* = \exp\left[-TE_a(R) + 0(T)\right].$$

The exponent $E_p(R)$ is estimated in Appendix D and may be written

$$E_p(R) = WE(R/W),$$

where $E(\hat{R})$ is defined by (111). The exponent $E_a(R)$ is estimated in Refs. 10, 8, 9, 13. The exponents are compared in Fig. 8 for $A = 2$ by plotting their known lower bounds. It is also possible to show that $E_a(0) = E_p(0) = AW/2$ for all A.

REFERENCES

1. Viterbi, A. J., Systematic Coding for the Continuous Gaussian Channel, Ph.D. Dissertation, University of Southern California, August, 1962.
2. Balakrishnan, A. V., A Contribution to the Sphere-Packing Problem of Communication Theory, J. Math. Anal. Appl., *3*, 1961, pp. 485–506.
3. Blichfeldt, H. F., The Minimum Value of Quadratic Forms, and the Closest Packing of Spheres, Math Ann., *101*, 1929, pp. 605–608.
4. Watson, G. N., *The Theory of Bessel Functions*, Cambridge Univ. Press, London, 1958.
5. Erdélyi, A., et al., *Higher Transcendental Functions*, Vol. 2, McGraw-Hill Book Company, New York, 1953.
6. Grobner, W. and Hofreiter, N., *Integraltafel* (*Vol. II*), Springer-Verlag, Vienna, 1958.
7. Chernoff, H., A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations, Ann. Math. Statist., *23*, 1952, pp. 493–507.
8. Gallager, R. G., A Simple Derivation of the Coding Theorem and Some Applications, IEEE Trans. Inform. Theor., *IT-11*, January 1965, pp. 3–18.
9. Shannon, C. E., Gallager, R. G., and Berlekamp, E. R., forthcoming paper to be published in Inform. and Control.
10. Shannon, C. E., Probability of Error for Optimal Codes in the Gaussian Channel, B.S.T.J., *38*, May, 1959, pp. 611–656.
11. Nuttal, A. H., Error Probabilities for Equicorrelated M-ary Signals Under Phase-Coherent and Phase-Incoherent Reception, IEEE Trans. Inform Theor., *IT-8*, July, 1962, pp. 305–314.
12. Shannon, C. E., Communication in the Presence of Noise, Proc. IRE, *37*, Jan., 1949, pp. 10–21.
13. Wyner, A. D., An Inproved Error Bound for Gaussian Channels, B.S.T.J., *43*, Nov., 1964, pp. 3070–3075.

Synchronization Recovery Techniques for Binary Cyclic Codes*

By S. Y. TONG

(Manuscript received December 17, 1965)

A class of binary block codes capable of simultaneous correction of additive errors and synchronization errors is presented. This class of codes consists of coset codes of binary cyclic or shortened cyclic codes, and retains the implementation advantages of binary cyclic codes. In most cases, the redundancy required to provide additive-error correction is sufficient to give synchronization-error correction so that no additional redundant bits are required.

Synthesis procedures to construct such codes are also presented, along with an upper bound on the number of synchronization errors which can be corrected by codes in this class.

I. INTRODUCTION

In serial-type data transmission systems, alpha-numeric characters are ordinarily represented by groups of binary symbols. To get meaningful information transfer, it is necessary at the receiver to correctly partition the incoming bit sequence, i.e., to establish and maintain "Character Timing". It is well known that channel noise not only produces additive errors but also can cause timing errors; consequently, methods to correct timing errors have been suggested by many authors. Usually these methods require special coding of the messages, as in comma-free codes,[1,2,3,4,5] or the insertion of synchronization sequences between blocks of messages.[6]

A similar timing problem exists in systems where error control is employed; the problem is transformed from character timing to word synchronization, or, equivalently, the ability to distinguish information bits from check bits. Codes that protect word synchronization as well

[*] This is a part of the Ph.D. dissertation submitted by the author to the Department of Electrical Engineering, Princeton University, Princeton, N. J.

as correct additive errors have been investigated. Sellers[7] has proposed a scheme where a burst-error-correcting code interlaced with additional check bits is used to give limited protection against synchronization loss, or provide burst-error correction when synchronization is maintained. Stiffler[8] has derived a necessary and sufficient condition for the existence of coset codes such that the sequences produced by slipping the word framing by r bits will not be code words. Such coset codes are useful for systems where coding is used for error detection only.

In order to utilize such a coset code for both additive-error correction and synchronization-error detection it is necessary that synchronization errors not result in decodable sequences. A condition sufficient for this has been obtained by Levy.[9] However, a more useful result would be a condition which would enable *correction* of both types of errors.

In this paper, a technique for obtaining codes capable of correcting synchronization errors as well as additive errors is presented. Furthermore, it is shown that in many cases correction of synchronization errors is possible even in the presence of additive noise. Generally, the scheme requires no additional check bits and the implementation is simple.

In addition to these fundamental results, a set of coset codes, optimal with respect to synchronization error detecting ability is obtained; this represents an improvement of Levy's results.[9]

1.1 *Definitions and Preliminaries*

To correct word-sync loss for an error-correcting code there are two conditions which must be met in order not to reduce the normal error-correcting capability of the code. The first condition is that an error pattern caused by sync loss must not be in any of the cosets which the code utilizes for correction of additive errors. This will ensure that the loss of sync will not be interpreted by the decoder as additive noise, and vice versa. The second condition is that the set of error patterns caused by misframing in one direction must be disjoint from the set of error patterns caused by misframing in the other direction. If the second condition is met, one can proceed to correct the word framing error iteratively. For the detection of sync loss, the first condition is necessary and sufficient, while for the correction of sync loss, the second condition must be satisfied. If the code is used for detection only, the first condition reduces to the requirement that the overlapping of any two code words should not be another code word, so that sync errors can always be detected as an erroneous word, although one would not be able to distinguish sync loss from additive errors.

Let

$$\mathbf{A} = (a_1, a_2, \cdots, a_n)$$
$$\mathbf{B} = (b_1, b_2, \cdots, b_n)$$
$$\mathbf{C} = (c_1, c_2, \cdots, c_n)$$
$$\mathbf{D} = (d_1, d_2, \cdots, d_n)$$

be code words, not necessarily distinct, then:

Definition 1: A synchronization bit loss (or bit loss) of r bits in word framing is said to occur if the receiver bit counter is r bits behind what it should be. That is to say, if the sequence $a_1, \cdots, a_n, b_1, \cdots, b_n$ is framed by the receiver as $a_{n-r+1}, \cdots, a_n b_1, \cdots, b_{n-r}$ where $r < [n/2]$ and the message is taken in such a way that b_n is the first bit of the sequence to arrive at the receiver.

Definition 2: A synchronization bit gain (or bit gain) of r bits in word framing is said to occur if the receiver bit counter counts r bits more than it should. Thus, $a_1, \cdots, a_n, b_1, \cdots, b_n$ is framed by the receiver as $a_{r+1}, \cdots, a_n b_1, \cdots, b_r$ where $r \leqq [n/2]$.

Definition 3: If $(a_{n-i+1}, \cdots, a_n b_1, \cdots, b_{n-i})$ and $(a_{i+1} a_{i+2}, \cdots, a_n b_1 b_2, \cdots, b_i)$ are not code words for every pair, **A**, **B** and all i,

$$0 < i \leqq r,$$

then the code is said to have comma-free freedom r.

Definition 4: A correctable coset of a code is defined as a coset whose leader is one of the error patterns the decoder corrects.

Definition 5: If the sequences $a_{i+1} a_{i+2}, \cdots, a_n b_1 b_2, \cdots, b_i$ and $a_{n-i+1}, \cdots, a_n b_1, \cdots, b_{n-i}$ do not belong to any of the correctable cosets of the code for every pair **A**, **B** and all i, $0 < i \leqq r$, then the code is said to have sync-detection capability r.

Definition 6: A code is said to have sync-recovery capability r if the code has sync-detection capability $q \geqq r$ and if the cosets containing $a_{i+1}, \cdots, a_n b_1, \cdots, b_i$ are disjoint from the cosets containing $c_{n-j+1}, \cdots, c_n d_1, \cdots, d_{n-j}$ for all $0 < i, j \leqq r$ and for every set of **A**, **B**, **C**, **D** of the code.

Definition 7: A code is said to have guaranteed noise tolerance of (k,r) bits, if the code is guaranteed to correct r bits of sync slippage with k or fewer additional bit errors in the received block.

Definition 8: A code is said to have expected noise tolerance of $E(r)$ bits, if with a probability of at least $\frac{1}{2}$, the code can correct r bits of sync slippage with $E(r)$ or fewer additional bit errors in the received block.

In the subsequent discussion all the error-correcting codes are assumed to correct random errors.

1.2 *Summary of the Results*

Consider a binary cyclic code which corrects $t \geqq 2r + 1$ errors, $r > 0$. If such a code is shortened by $2r + 1$ bits or more, it is shown in Section II that the code can be made to have sync-recovery capability r with expected noise tolerance of $E(\beta)$ bits when a slippage of β bits occurs

$$E(\beta) = 2^{-2\beta} \sum_{i=1}^{2\beta} (t - i) \binom{2\beta}{i - 1}, \qquad 0 < \beta \leqq r$$

without added redundancy. A similar technique is developed for codes which correct more than one error. It is shown that by adding $2r$ zeros to each code word and shortening the code by $2r + 1$ bits or more that the code can be made to have sync-recovery capability of r bits.

A scheme which is applicable to a single error-correcting code is also developed. It is shown that for any error-correcting code without an even-parity check it is possible to have sync-recovery capability of one bit if two information bits of the code are replaced by two zeros. In Section III, techniques are developed for binary cyclic codes which are not shortened. A necessary and sufficient condition is derived for the existence of a coset code having specified sync-correcting ability. It is also shown that a cyclic code which corrects t errors and has minimum distance d_m can be made to have sync-recovery capability

$$r \leqq d_m - 2t - 2$$

without additional redundancy and an optimal set of codes that assures $r = d_m - 2t - 2$ is given.

For cyclic codes with some special properties, it is shown that a synthesis procedure can be used to construct coset codes of the given code so that one bit out-of-sync can always be corrected. This procedure applies to almost all of the Bose-Chaudhuri[10] Hocquenghem[11] codes that corrects more than two errors.

Bounds on the amount of slippage which can be corrected are derived. It is shown that sync-recovery capability for any binary cyclic code cannot exceed $[(n - k - 1)/2]$* bits and sync-detection capability cannot

* $[x]$ denotes the greatest integer less than or equal to x.

exceed $n - k - 1$ bits, where n is the block length of the code and k is the number of information bits in the code. A brief discussion on the sync-detection capability of error-detecting codes is included in Section IV.

1.3 *Properties of Cyclic Codes*

A subspace V of l-tuples is called a cyclic code if for each vector $v = (a_0, a_1, \cdots, a_{l-1})$ in V, the vector $v' = (a_{l-1}, a_0, \cdots, a_{l-2})$ is also in V.

By considering each l-tuple as an element of the algebra A_l of polynomials modulo $x^l = 1$, one may associate each l-tuple $(a_0, \cdots, a_{l-1})$ with a polynomial $f(x) = a_0 + a_1x, \cdots, a_{l-1}x^{l-1}$ in the residue class modulo $x^l - 1$.* It can be shown that a subspace is a cyclic code if and only if it is an ideal in the algebra of polynomials modulo $x^l - 1$.[12] The generator $g(x)$ of the ideal is known as the generator polynomial of the cyclic code. It follows that l, which represents the natural length of the code, must be the least common multiple of the roots of the generator polynomial of the cyclic code. Given an (l,k) cyclic code, it is always possible to form an $(l - i,k - i)$ code by making the i leading information bits identically zero and omitting them from all code vectors. Such a code is no longer cyclic, and is called a shortened cyclic code. Denote the code space of a cyclic code by C_0 and the code space of a shortened cyclic code by C_i, where i is the number of bits shortened. Let n be the block length of a shortened cyclic code (i.e., $n = l - i$), the higher order $l - n$ bits are identically zero and hence are not transmitted. We can imagine that the receiver will decode the shortened code by first augmenting the received n-bit code word by $l - n$ zeros and then decoding it as if it were a full-length cyclic code. (Note that the actual receiver need not perform these precise functions, but in any case it has to do something equivalent to this.)

Now let us investigate what will happen if an r-bit loss occurs as shown in Fig. 1. Given that the transmitted message is $R(x)$, the received message is

$$x^r R(x) + A(x) + x^n B(x) \tag{1}$$

where $A(x)$ is the portion of next word entered into the framing and $B(x)$ is the higher order r-bit portion of $R(x)$ out of framing. In general,

* This paper discusses codes over binary field only so that the coefficients of polynomials are either 0 or 1 and + or − signs are used interchangeably.

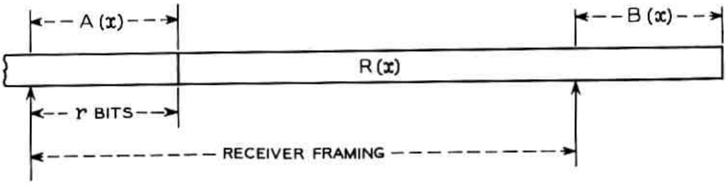


Fig. 1 — The situation of synchronization loss.

$A(x)$ and $B(x)$ are not predictable and one may write the received message in the following form

$$x^r R(x) + \delta_1{}^r(x) + x^n \delta_2{}^r(x) \tag{2}$$

where $\delta_i{}^r(x)$ represents a polynomial of degree at most $r - 1$ with random coefficients. For cyclic codes, $n = l$ and $x^l = 1$. Therefore,

$$\delta_1{}^r(x) + x^l \delta_2{}^r(x) = \delta_1{}^r(x) + \delta_2{}^r(x) = \delta^r(x), \tag{3}$$

so that the received message is

$$x^r R(x) + \delta^r(x). \tag{4}$$

Since $x^r R(x)$ is an element of C_0, it is divisible by $g(x)$ so that its syndrome is 0. The coset leader of the coset to which $\delta^r(x)$ belongs must have a weight of no more than r.* For the same reason, with a shortened cyclic code the coset leader of the coset of $\delta'(x) = [\delta_1{}^r(x) + x^n \delta_2{}^r(x)]$ must have a weight of not more than $2r$. It can be shown that a similar result holds for an r-bit gain. It follows that:

Theorem 1: A slippage of r bits in synchronization can result in a vector at most distance $2r$ from a nearest code vector if the code is a shortened cyclic code and at most distance r from a nearest code vector if the code is cyclic, where code vectors are elements of C_0.

II. SCHEME FOR SHORTENED CYCLIC CODES

2.1 *Codes that Correct at Least Three Random Errors*

Suppose a fixed polynomial, $P(x)$ is added to every code word at the encoder and the same polynomial is subtracted from the received message at the decoder. If the sync is maintained properly, the effect of $P(x)$ will be canceled. However, if sync is not maintained, an error

* We assume that the coset leader is the minimum weight element of the coset, which is the desired case for a random-error correcting code.

pattern may be detected. By choosing $P(x)$ in a suitable way, it is conceivable that one may be able to detect or even correct sync slips.

Let the code word be $R(x) \epsilon C_i = C_{l-n}$, where $x^l = 1$ and n is the length of shortened code word. The transmitted word is $R(x) + P(x)$. At the receiver, assume the word framing has an r-bit loss as shown in Fig. 1. The received word, $Q(x)$, according to (2), takes the following form:

$$Q(x) = x^r[R(x) + P(x)] + \delta_1^r(x) + x^n\delta_2^r(x). \qquad (5)$$

The receiver then subtracts $P(x)$ from $Q(x)$, i.e.,

$$Q_1(x) = Q(x) + P(x). \qquad (6)$$

The syndrome of $Q_1(x)$ is the remainder of $Q_1(x)/g(x)$, i.e.,

$$\begin{aligned} \{Q_1(x)\} &= \{Q(x) + P(x)\} \\ &= \{(1 + x^r)P(x) + \delta_1^r(x) + x^n\delta_2^r(x)\} \end{aligned} \qquad (7)$$

where $\{x\}$ represents either the residue class of x modulo $g(x)$ or the polynomial of at least degree in that class; the context should make it clear which is meant. If $P(x)$ is chosen so that the coset $\{(1 + x^r)P(x)\}$ coincides with $\{x^i + \theta_1^r(x) + x^n\theta_2^r(x)\}$, where $n \leqq i < l$, and $\theta_j^r(x)$ is a polynomial of degree at most $r - 1$, for $j = 1, 2$, then,

$$\begin{aligned} &\{(1 + x^r)P(x) + \delta_1^r(x) + x^n\delta_2^r(x)\} \\ &\qquad = \{x^i + \theta_1^r(x) + x^n\theta_2^r(x) + \delta_1^r(x) + x^n\delta_2^r(x)\} \qquad (8) \\ &\qquad = \{x^i + \delta_3^r(x) + x^n\delta_4^r(x)\}. \end{aligned}$$

Note that $x^i + \delta_3^r(x) + x^n\delta_4^r(x)$ has at most $2r + 1$ nonzero terms. Thus, if the code C_0 corrects $t \geqq 2r + 1$ errors, the polynomial

$$x^i + \delta_3^r(x) + x^n\delta_4^r(x)$$

must be a coset leader of C_0 for all possible $\delta^r(x)$'s. Thus, if $n + r \leqq i$, x^i cannot be canceled by the $\delta^r(x)$'s and the decoder will indicate that the $i + 1$ position of the received word is in error. But $n \leqq i < l$, so the ith $+ 1$ bit was not transmitted; this can be used to indicate that a misframing has occurred.

Similarly, it can be shown that if the receiver has an r-bit gain, then the message to the decoder is

$$Q_2(x) = x^{-r}[R(x) + P(x)] + x^{-r}\delta_1^r(x) + x^{n-r}\delta_2^r(x) + P(x) \qquad (9)$$

$$\begin{aligned} \{Q_2(x)\} &= \{x^{i-r} + x^{-r}\delta_3^r(x) + x^{n-r}\delta_4^r(x)\} \\ &= \{x^{-r}Q_1(x)\} \qquad \text{for some } Q_1(x). \end{aligned} \qquad (10)$$

Since each possible $Q_1(x)$ is a coset leader for all $\delta^r(x)$, so is $x^{-r}Q_1(x)$. Thus, by reasoning similar to that employed in the bit loss case, an r-bit gain can be detected.

In order to recover synchronization, one must be able to distinguish the syndrome due to bit loss from the syndrome due to bit gain. This implies:

$$\{x^i + \delta_3^r(x) + x^n\delta_4^r(x)\} \neq \{x^{i-r} + x^{-r}\delta_3^r(x) + x^{n-r}\delta_4^r(x)\} . \qquad (11)$$

That is, the error patterns must not be in the same coset.

$$\{x^i + x^{i-r} + \delta_3^r(x) + x^n\delta_4^r(x) + x^{-r}\delta_3^r(x) + x^{n-r}\delta_4^r(x)\} \neq 0. \qquad (12)$$

The weight of the polynomial within the bracket is at most $4r + 2$. Since the code is $t \geqq 2r + 1$ error correcting, any code word must have a weight of at least $4r + 3$. Hence, the inequality is always satisfied provided that either x^i or x^{i-r} is not canceled by the terms of the $\delta_i(x)$ polynomials. Otherwise, it would be possible for all other terms of $\delta_i(x)$ to be zero, resulting in a zero polynomial and thereby contradicting (12).

It is clear that in order to do this it is necessary and sufficient to have either x^i or x^{i-r} or both not in those positions where the δ's may take the value of 1's; that is to say, either

$$n + r - 1 < i < l - r \qquad (13)$$

or

$$n + r - 1 < i - r < l - r \qquad (14)$$

must be satisfied. Both conditions require $l - n \geqq 2r + 1$ which is the minimum number of bits required to be eliminated.

Thus, we have shown that if: (*i*) a cyclic code corrects $t \geqq 2r + 1$ errors; (*ii*) the number of eliminated bits is at least $2r + 1$; and (*iii*) it is possible to find a polynomial $P(x)$ constrained by (8); then,upon an r-bit loss or gain, the syndrome generated will be different from that of any of the correctable cosets and every bit loss syndrome will be different from any of the syndromes caused by bit gain, and vice versa. To complete the analysis, one must show the existence of polynomials $P(x)$ constrained by (8).

Recall that the requirement on $P(x)$ is that

$$\{(1 + x^r)P(x)\} = \{x^i + \theta_1^r(x) + x^n\theta_2^r(x)\} . \qquad (15)$$

Since the θ's are arbitrary polynomials, [as they will be combined with δ's, see (8)] we may treat them as variables in determining the simplest possible $P(x)$. In particular, if one elects to satisfy condition (13), i.e.

$$n + r - 1 < i < l - r$$

and to minimize the number of bits to be eliminated, the smallest i should be selected. Therefore, let $i = n + r$. For reasons which will become apparent later,* we set $\theta_1^r(x) = 0$ and $\theta_2^r(x) = 1$. Then

$$\{(1 + x^r)P(x)\} = \{x^{n+r} + x^n\}$$

which implies

$$\{P(x)\} = \{x^n\}. \tag{16}$$

Thus, we have constructed a $P(x)$ which satisfies (15). This completes the derivation of $P(x)$ for the detection and correction of an r-bit gain or loss. To see if this pattern is also good for any β-bit sync slippage $(0 < \beta \leqq r)$, we note that the error pattern for a β-bit loss is

$$\begin{aligned}(1 + x^\beta)P(x) + \delta_1^\beta(x) + x^n\delta_2^\beta(x) \\ &= x^n + x^{n+\beta} + \delta_1^\beta(x) + x^n\delta_2^\beta(x) \qquad (17) \\ &= x^{n+\beta} + \delta_1^\beta(x) + x^n\delta_3^\beta(x).\end{aligned}$$

The number of nonzero terms of the polynomial is at most

$$2\beta + 1 < 2r + 1 \leqq t, \qquad \text{for all} \qquad \beta < r,$$

hence, identification of an error in position $n + \beta + 1$ is always possible. Since $n < n + \beta + 1 < l$, and position $n + \beta + 1$ was not transmitted, the fact that the decoder will show the $n + \beta + 1$ position to be in error can be used to indicate that an out-of-sync situation exists. By a similar argument, it can be shown that the detection of any β-bit gain is also possible for $\beta < r$.

In order to recover synchronization, the set of syndromes corresponding to bit loss must be different from those corresponding to bit gain; that is,

$$\begin{aligned}\{x^{n+\alpha} + \delta_1^\alpha(x) + x^n\delta_3^\alpha(x)\} \\ &\neq \{x^n + x^{-\beta}\delta_1^\beta(x) + x^{n-\beta}\delta_3^\beta(x)\}\end{aligned} \tag{18}$$

for all

$$0 < \alpha, \beta \leqq r$$

or

$$\{x^{n+\alpha} + x^n + \delta_1^\alpha(x) + x^n\delta_3^\alpha(x) + x^{-\beta}\delta_1^\beta(x) + x^{n-\beta}\delta_3^\beta(x)\} \neq 0. \tag{19}$$

* To maximize expected noise tolerance, it is desirable to keep the degree of $\theta^r(x)$ small.

The number of nonzero terms within the brackets (19) is at most

$$2\alpha + 2\beta + 2 < 4r + 3 < 2t + 1 \leqq d_m$$

where d_m is the minimum distance of the code. Equation (19) cannot represent a code word unless both $x^{n+\alpha}$ and x^n are canceled by some terms of the δ's in such a way that all an-zero polynomial results. It is seen from (19) that $n + \alpha < l - \beta$ is a sufficient condition for $x^{n+\alpha}$ not to be canceled by any of the δ's. Since $l - n \geqq 2r + 1 > \alpha + \beta$, this is always satisfied. Thus, we have shown that a shortened cyclic code which corrects t random errors will have sync-recovery capability of r bits, if $l - n \geqq 2r + 1, t \geqq 2r + 1$, and if the *syndrome* of x^n is added to each word after encoding and before decoding.

The implementation of this scheme is obviously easy. The generation of $\{x^n\}$ can be accomplished by adding one bit corresponding to position x^n at encoder and decoder. (Since this position is not actually transmitted over the channel, only the syndrome of x^n is added to the encoded word.) Usually only a slight loss in efficiency results from the shortening of the code.

The decision rule can be formed as follows:

(*i*) If the decoder indicates that the bit corresponding to x^n is in error, but not any x^k, $n < k \leqq l - 1$, then one assumes the system has gained a few bits in bit count. By extending the sync count one bit at a time, the system will regain sync in at most r word times.

(*ii*) If the decoder indicates that the bit corresponding to $x^{n+\beta}$ is in error, $1 \leqq \beta \leqq r$, then one assumes the system has lost β bits. The word sync can be recovered either by using step-by-step correction as in (*i*), or by a one step correction.

Note that the $P(x)$ derived here is but one of many possibilities; for example, by letting $\{P(x)\} = \{x^{l-1}\}$ one would come up with a similar decision rule which favors bit-gain correction rather than the bit-loss correction as we have done.

Example: Consider the (23,12) Golay code shortened to a (20,9) code. Since it is a triple error-correcting code and since $l - n = 3$, it should have sync-recovery capability of 1 bit. Let us demonstrate this fact by step-by-step computation. The generator polynomial for the code is:

$$g(x) = 1 + x^2 + x^4 + x^5 + x^6 + x^{10} + x^{11}$$

$$= 101011100011.$$

The syndrome of $x^n = x^{20}$ is 00101101111 $= \{P(x)\}$ and of $x^{n+1} = x^{21}$ is 10111000110. Assume the information 000000001 is to be transmitted. After encoding, the code word is

$$\begin{array}{rl} & 1{,}0101101111000000001{,}01 \\ \text{Add } P(x) & \underline{0010110111100000000} \\ & 1{,}0111011000100000001{,}01. \end{array}$$

The above sequence is the transmitted message.

(*i*) Assume one bit loss has occurred. Thus, the received message is

$$\underset{\uparrow}{1},01110110001000000\underset{\uparrow}{0}1,01. \qquad \overset{\text{message flow}}{\longrightarrow}$$

Receiver Frame

To this word the receiver adds $P(x)$ and three 0's

$$\begin{array}{rl} & 1011101100010000000 \\ & \underline{0010110111100000000} \\ A(x) = & 1001011011110000000\,(000) \end{array}$$

at the high-order end, as shown in the parentheses, to make a cyclic code. The syndrome of the message, $A(x)$, is the remainder of $A(x)/g(x)$.

$$\{A(x)/g(x)\} = 00111000110$$

$$= \{x^{21}\} + \{x^0\}.$$

d x^0 are in
3y the deci-
urred.
ed message

x^{19}.

The decoder will indicate that bits corresponding to x^{21} an
error, but x^{21} was not transmitted and $x^{21} = x^{20+1} = x^{n+\beta}$. E
sion rule just derived, one decides that a one bit loss has occ

(*ii*) Suppose a one bit gain has occurred. Thus, the receiv
is

$$\underset{\uparrow}{1},01110110001000000\underset{\uparrow}{0}1,01;$$

add $P(x)$ and three missing zeros. We have

$$\begin{array}{rl} & 1110110001000000010 \\ P(x) = & \underline{0010110111100000000} \\ & 1100000110100000010\,(000). \end{array}$$

The syndrome is

$$01110110001 = 00101101111 + 01011011110 = x^{20} +$$

By the decoding rule we see that $x^n = x^{20}$ is in error but not x^k for all k such that $20 = n < k \leqq l - 1 = 22$, hence the decoder decides that a bit gain has occurred. Notice that in this case, only two errors are indicated by the decoder due to a misframing of one bit. Because the code is triple error-correcting, sync recovery of a one bit of misframe is possible even if there is an error due to additive noise. This feature of noise tolerance will be discussed in the following paragraph.

2.1.1 *Noise Tolerance*

Assume a slippage of $\beta \leqq r$ bits has occurred. According to (17) the error pattern is

$$x^{n+\beta} + \delta_1^{\beta}(x) + x^n\delta_3^{\beta}(x).$$

The weight of this error polynomial is at most $2\beta + 1$, but the code corrects t errors, so that at least $t - 2\beta - 1$ additional errors can be in the received block without disturbing the sync-correcting process. This is so because every error-bit position can be identified correctly provided the total number of errors does not exceed t; hence the guaranteed noise tolerance is $(t - 2\beta - 1,\beta)$.

Since channel noise cannot affect any of the bit positions $n < i \leqq l$ directly, the vital bits for detection and correction of sync are not likely to be corrupted by noise. In fact, it can be shown that at least $2r + 2$ additional errors are required at specific locations to change the bit in position i for $n < i \leqq l$.

If we assume that the probability of occurrence of any nonzero term of the δ's is 1/2, we can compute the expected noise tolerance $E(\beta)$ as follows:

If β bits of slippage occurs, there is at least one error caused by the error pattern purposely generated by $P(x)$, but not more than $2\beta + 1$ errors, in accordance with (17). The probability of occurrence of a total of i errors due to sync slippage of β bits is

$$P_i = \binom{2\beta}{i-1} \Big/ 2^{2\beta}; \tag{20}$$

the number of additional errors which can be tolerated with i errors is $t - i$, so the expectation is

$$E(\beta) = 2^{-2\beta} \sum_{i=1}^{2\beta+1} (t - i) \binom{2\beta}{i-1}. \tag{21}$$

That is to say, on the average, for a β-bit sync slippage, one can tolerate $E(\beta)$ additional errors. For example, consider a (255,215) BCH code which corrects five random errors. Since $t \geqq 2r + 1$, $r_{\max} = 2$, $l - n \geqq 2r + 1 = 5$, $l = 255$, $n = 250$; therefore, the (250,210) shortened cyclic code can have sync-recovery capability up to 2 bits per word.

The guaranteed noise tolerances are $(t - 2\beta - 1, \beta) = (2,1)$ and $(0,2)$, and the expected noise tolerances are

$$E(1) = 2^{-2} \sum_{i=1}^{2} (5 - i) \binom{2}{i - 1} = 3$$

$$E(2) = 2^{-4} \sum_{i=1}^{4} (5 - i) \binom{4}{i - 1} = 2.$$

2.2 *Double Error-Correcting Codes*

The scheme just proposed cannot be used for double error-correcting codes because, in the worst case, for only one bit of sync slippage, it is possible to have three errors generated as a result of sync loss. (However, the probability of recovering sync loss is still high.) In this section the scheme is modified so that it is guaranteed to correct sync loss for double error-correcting codes; however, extra redundancy is required.

Let m zeros be placed on each end of a coded message of an n-bit shortened cyclic code. The transmitted word is then a stream of binary messages interlaced with $2m$-zero bits between messages (i.e., the added zeros are actually transmitted). A typical word is of the form shown below

$$\begin{array}{ccc} x^0 x^1 \cdots x^{m-1} & x^m \cdots x^{m+n-1} & x^{m+n} \cdots x^{2m+n-1} \\ \text{all zero} & x^m R(x) & \text{all zero} \end{array}$$

where the first and the last m bits are identically zero and the center portion is the shortened code word, $R(x)$.

Let $x^m P(x)$ be added to such a code; the transmitted message is

$$x^m [R(x) + P(x)]. \tag{22}$$

It is clear that the word framing at the receiver must contain $2m + n$ bits. It follows that all terms of $x^{r+m}[R(x) + P(x)]$ will be within a single receiver frame for all $|r| \leqq m$. Thus, for an r-bit loss the received message, in the absence of noise, is

$$x^{r+m}[R(x) + P(x)].$$

After subtracting $x^m P(x)$ at receiver, the syndrome of the resultant message is

$$\{x^{r+m}[R(x) + P(x)] + x^m P(x)\} \\ = \{x^m(1 + x^r)P(x)\}, \qquad |r| \leqq m. \tag{23}$$

Let $P(x) = \{x^{l-m-1}\}$, then the syndrome due to an r-bit loss is

$$\{x^m(1 + x^r)P(x)\} = \{x^{l-1} + x^{l+r-1}\}. \tag{24}$$

If $2m + n \leqq l - 1$, the bit corresponding to x^{l-1} (which is the lth bit) is not transmitted. Hence, the detection of x^{l-1} in error can be used to indicate a bit loss. In order to correct the bit loss, note that the error at x^{l+r-1} corresponds to the bit at position $l + r$ modulo l. Therefore, if $r > 0$, the error bit position corresponds to one of the first m bits which is known to be zero so that the inconsistency between the calculated error bit and the actual bit value at position r can serve as an indication of an r-bit loss.

Similarly, for $r < 0$ the bit corresponding to the syndrome $\{x^{l+r-1}\}$ is $l + r \geqq 2m + n + r + 1 \geqq m + n + 1$ for $0 > r \geqq -m$ but the bits $m + n + 1, m + n + 2, \cdots, 2m + n$ are known to be zero, so that the r-bit gain can be detected in the same way. Notice that if $l + r \geqq 2m + n$, for some $r < 0$, the error indication directly shows an $|r|$-bit gain has occurred since the bit $l + r$ is not transmitted.

Recall that one requires

$$2m + n \leqq l - 1, \tag{25}$$

i.e.,

$$l - n \leqq 2m + 1 \tag{26}$$

so that at least $2m + 1$ bits of the information symbols must be eliminated.

Since exactly two errors (namely x^{l-1} and x^{l+r-1}) will be generated for every sync loss up to m bits, a code capable of correcting at least two errors must be used. It is obvious that for a t-error-correcting code, an addition of $t - 2$ errors anywhere in the data section (i.e., in the $x^m R(x)$ part) can be tolerated. However, errors in the zero section can affect the sync-correcting process, although the sync-detection capability will not suffer. Thus, the guaranteed noise tolerance for sync-recovery is zero while it is $t - 2$ for sync detection.

We summarize the sync-correcting rule as follows:

(*i*) If x^{l-1} is in error then one assumes that a sync-slip has occurred.

(*ii*) If, in addition, one and only one of the calculated error bits is in the bit position corresponding to x^{l+r-1}, $|r| \leqq m$, while the actual bit of that position is either not transmitted or has the value of zero then one assumes an r-bit loss has occurred (if $r < 0$, $| r |$-bit loss $= r$-bit gain).

It is easy to see from the above argument that any $P(x) = \{x^{l-i}\}$, $1 \leqq i \leqq l - 2m - n$, can be used for this scheme.

Example: Consider a (15,7) BCH code shortened to (12,4) code with one zero adjoined at each end of the code to obtain (14,4) code. Here $l = 15$, $m = 1$, $n = 12$. The code is generated by

$$g(x) = 1 + x^4 + x^6 + x^7 + x^8.$$

Assume the message 1100 is to be sent. After encoding we have the (12,4) code word

$$\begin{aligned} R(x) &= 010001011100 \\ P(x) &= \{x^{l-m-1}\} = \{x^{13}\} = 001011100000 \\ R(x) + P(x) &= 011010111100. \end{aligned}$$

Add a zero at each end, the transmitted message is

$$0{,}00110101111000{,}0.$$

Notice that the neighboring bits of the message must be zero since each message must begin and end with zero, by construction.

(*i*) At the receiver assume a one-bit loss has occurred. Then we have the message, as the receiver sees it:

$$\underset{\uparrow}{0}{,}0011010111100\underset{\uparrow}{0}{,}0.$$

The receiver adds $xP(x)$ and the resulting message becomes

$$\begin{array}{rl} & 00011010111100 \\ \text{add } xP(x) = & 00010111000000 \\ \hline & 00001101111100(0). \end{array}$$

The syndrome is $10010111 = 00010111 + 10000000 = \{x^{14}\} + \{x^0\}$ which indicates x^{14} and x^0 are in error but x^{14} is never transmitted and x^0 is known to have been transmitted as zero. Therefore, according to the decoding rule just derived, we see that word sync has slipped. Since $x^{l+r-1} = x^{14+r} = x^0$, $r = 1$, which shows a gain of one bit has occurred.

(*ii*) Assume one-bit gain has occurred, the received message becomes as shown

$$\begin{array}{rl} & 01101011110000 \\ \text{add } P(x) = & 00010111000000 \\ \hline & 01111100110000(0). \end{array}$$

The syndrome is 00111001 = 00010111 + 00101110 = $\{x^{14}\} + \{x^{13}\}$. x^{14} in error indicates a sync-slip condition and $x^{13} = 0$

$$x^{l+r-1} = x^{13} = x^{15-1-1},$$

so $r = -1$ which shows that a gain of one bit in word sync has occurred.

2.3 *Codes Without an Even-Parity Check*

Recall that, from (23), the syndrome is $\{x^m(1 + x^r)P(x)\}$ for an r-bit loss. It is desirable to have such a syndrome coincide with the syndrome of a single bit, say x^i *, and if the bit corresponding to x^i is either not transmitted, or if the value of the coefficient of x^i is known to the receiver by prearrangement, then it is possible to have sync recovery capability for a single error-correcting code. We shall show in the following that such a possibility does exist.

Assertion: For a cyclic error-correcting code without an even-parity check† it is always possible to modify the code in such a way as to have sync-recovery capability of one bit. The scheme is based on the following theorem.

Theorem 2: If $g(x)$ is a generator polynomial for a cyclic code of natural length l and if

$$\text{GCD }(g(x),1 + x) = 1$$

then

$$(1 + x) \mid \{x^{l-1}\}.$$

Proof:
$$(1 + x) \nmid g(x) \Rightarrow g(1) \neq 0$$

$$\therefore\ g(1) = 1 \quad \text{or} \quad 1 + g(1) = 0$$

or $1 + g(x)$ is divisible by $1 + x$.
Likewise $x \nmid g(x)$ because $g(x)h(x) = x^l + 1$ and if $x \mid g(x)$ then $x \mid x^l + 1$ which is impossible.

Therefore,

$$x \mid [1 + g(x)]. \tag{27}$$

It follows that $1 + g(x) = x(1 + x)F_3(x)$. In the ring of polynomials modulo $x^l + 1$

$$x^l \equiv 1$$

* i.e., the remainder of $x^i/g(x)$.

† Or, equivalently, the generator polynomial of the code is not divisible by $1 + x$.

or

$$x^{l-1} \equiv x^{-1},$$

by (27),

$$x^{-1} \equiv x^{-1}(1 + g(x)) \equiv (1 + x)F_3(x) \ [\text{mod } g(x)],$$

i.e.,

$$\{x^{l-1}\} = (1 + x)F_3(x).$$

Q.E.D.

Theorem 2 shows the existence of $P(x) = \{x^{l-1}/1 + x\}$ for such codes. Thus, for any cyclic code which corrects one or more errors and whose generator polynomial, $g(x)$, is not divisible by $1 + x$, one may shorten the code by 2 bits, append one zero at each end of the encoded message and add the pattern $xP(x) = x\{x^{l-1}/1 + x\}$. The configuration is shown in Fig. 2. Note that the added zeros are actually transmitted.

When the system is in synchronization, the framing of the receiver is as shown in Fig. 2. The receiver first adds $xP(x)$ and then decodes the whole word.

From (23), one-bit loss gives the syndrome

$$\{x(1 + x)P(x)\} = \{x^{l-1}x\} = \{x^0\}$$

and one-bit gain gives the syndrome $\{x^{-1}(1 + x)xP(x)\} = \{x^{l-1}\}$. Note that both x^0 and x^{l-1} are inserted zero bits; hence, the decoder can use the decision rule as shown in Table I.

Example: Consider the single error-correcting Hamming code generated by $x^4 + x + 1$. For this code $l = 15$. From Table II:

$$P(x) = \left\{\frac{x^{l-1}}{1 + x}\right\} = \{x^{10}\}.$$

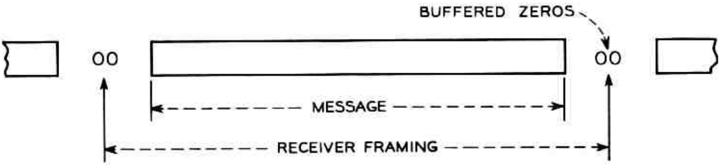


Fig. 2 — The relation between message and word framing.

TABLE I

Error Locator Indicates	Real Bit Value	Decision
x^0 in error	$x^0 = 0$	one-bit loss
x^0 in error	$x^0 = 1$	bit x^0 in error
x^{l-1} in error	$x^{l-1} = 0$	one-bit gain
x^{l-1} in error	$x^{l-1} = 1$	bit x^{l-1} in error

The code is modified from (15,11) to (15,9). Let the information bits be 001000101. After encoding it becomes 000100010001010.

Adding $xP(x) = x^{11}$, we have

$$\begin{array}{rl} xP(x) = & 000000000001000 \\ & \underline{000100010001010} \\ & 000100010000010. \end{array}$$

Case 1: One-bit loss:

The received message becomes

$$\begin{array}{rl} & 00001000100000100 \\ xP(x) = & \underline{00000000000100000} \\ & 00001000100100100. \end{array}$$

The syndrome is 1000; it indicates the first bit in error since the first bit is 0. From Table I we read "one-bit loss has occurred."

Case 2: One-bit gain:

The received message becomes

$$\begin{array}{rl} & 0010001000001000 \\ xP(x) = & \underline{0000000000010000} \\ & 0010001000011000. \end{array}$$

The syndrome is 1001; it indicates x^{14} in error but $x^{14} = 0$. So again by Table I we have detected one-bit gain.

TABLE II—GF(2^4)

x^0	1000	x^8	1010
x^1	0100	x^9	0101
x^2	0010	x^{10}	1110
x^3	0001	x^{11}	0111
x^4	1100	x^{12}	1111
x^5	0110	x^{13}	1011
x^6	0011	x^{14}	1001
x^7	1101	$x^{15} = x^0 =$	1000

2.3.1 *Noise immunity*

It is obvious that this scheme uses only one bit to detect and correct sync loss so that for codes which correct t errors we can afford $t - 1$ more additive errors in the code word provided that the bits x^0 and x^{l-1} are not in error.

III. FULL-LENGTH CYCLIC CODES

For cyclic codes, the technique used to distinguish synchronization loss from additive errors for shortened codes no longer applies. One must select $P(x)$ in such a way that over the range of synchronization slippage r, the error patterns so generated do not fall into any of the correctable cosets. This means

$$\delta^s(x) + P(x)(1 + x^s) \notin \mathcal{C} \tag{28}$$

and

$$x^{n-s}\delta^s(x) + P(x)(1 + x^{n-s}) \notin \mathcal{C}, \tag{29}$$

for all $\delta^s(x)$, $0 < s \leqq r$, where $\mathcal{C}$ is the set of all polynomials that belong to the union of all correctable cosets. By the following lemma it will be shown that conditions (28) and (29) are equivalent.

Lemma 1: $Q(x) \in \mathcal{C}$
if and only if

$$x^iQ(x) \in \mathcal{C} \qquad \textit{for all } i.$$

Proof: (*i*) Suppose $Q(x) \in \mathcal{C}$. Denote the coset leader to which $Q(x)$ belongs by $Q_c(x)$; then $Q(x) + Q_c(x) = w_Q(x)$ is a code word, but $x^iw_Q(x)$ is also a code word and the weight of $Q_c(x)$ is the same as the weight of $x^iQ_c(x)$; the weight of $Q_c(x)$ is no more than t, the maximum number of errors the code corrects. Therefore, $x^iQ_c(x)$ must also be a coset leader. It follows that $x^iQ_c(x) + x^iw_Q(x) = x^iQ(x)$ must belong to the coset whose leader is $x^iQ_c(x)$. That is to say $x^iQ(x) \in \mathcal{C}$.

(*ii*) $x^iQ(x) \in \mathcal{C} \Rightarrow x^{n-i}(x^iQ(x)) = Q(x) \in \mathcal{C}$ by (*i*). Q.E.D.

Since (29) can be rewritten as:

$$x^{n-s}[\delta^s(x) + P(x)(1 + x^s)] \notin \mathcal{C}, \tag{30}$$

by the above lemma and the law of contraposition, (28) and (29) are equivalent. The condition (28) can be restated as:

$$\bar{W}[Q_s(x)] \geqq t + 1 \tag{31}$$

where $Q_s(x)$ is the coset leader of the polynomial

$$\delta^s(x) + P(x)(1 + x^s)$$

and $\bar{W}[x]$ is defined as the weight of the polynomial x.

Equation (31) is a necessary and sufficient condition for a code to have sync-detection capability r. It is sufficient because if (31) is true, then no error pattern generated by $P(x)$ due to slippage of s bits,

$$0 < s \leqq r,$$

can be in a correctable coset; thus sync loss can be detected. It is necessary because otherwise there must exist at least one particular δ^s, say δ_α^s, such that (31) is not satisfied, then say

$$\bar{W}[Q_{s_\alpha}(x)] \leqq t, \qquad \text{for some } \alpha$$

where

$$Q_{s_\alpha}(x) = \delta_\alpha^s(x) + P(x)(1 + x^s), \qquad \text{for some } s, \qquad 0 < s \leqq r.$$

Then

$$\delta_\alpha^s(x) + P(x)(1 + x^s)$$

must be in one of the correctable cosets.

To have sync-correction capability r, the code must have sync-detection capability at least r and the syndromes of the error patterns for bit loss must be different than those for bit gain; that is

$$\{\delta^p(x) + P(x)(1 + x^p)\} \neq \{x^{n-q}[\delta^q(x) + P(x)(1 + x^q]\}, \quad 0 < p, q \leqq r \tag{32}$$

or

$$\{\delta^{p+q}(x) + P(x)(1 + x^{p+q})\} \neq 0 \tag{33}$$

or

$$\{\delta^{p+q}(x)\} \neq \{P(x)(1 + x^{p+q})\}. \tag{34}$$

If $p + q \geqq n - k$ then the degree of $\delta^{p+q}(x)$ can be as large as

$$p + q - 1 \geqq n - k - 1.$$

The right side of (34) has a degree at most $n - k - 1$. Thus, by proper choice of the coefficients of $\delta^{p+q}(x)$, one can always equate the two sides of (34). Hence, $p + q \leqq n - k - 1$ is a necessary condition to satisfy (34). But max p = max $q = r$ so that $2r \leqq n - k - 1$ is a necessary condition for the code to have sync-correction capability r. Hence the theorem:

Theorem 3: The sync-correction capability of a (n,k) coset code derived from a cyclic code cannot exceed $(n - k - 1)/2$.

If $p + q \leqq n - k - 1$, $\{\delta^{p+q}(x)\} = \delta^{p+q}(x)$ so that if the degree of $\{P(x)(1 + x^{p+q})\}$ is at least $p + q$, then (34) will be satisfied; *i.e.*,

$$D\{P(x)(1 + x^{p+q})\} \geqq p + q \tag{35}$$

is sufficient for (34) where $D[G(x)]$ denotes the degree of polynomial $G(x)$.

Therefore, the necessary and sufficient conditions for a code to have sync-correction capability r are:

$$\bar{W}[\delta^s(x) + P(x)(1 + x^s)] \geqq t + 1 \qquad \text{for all } 0 < s \leqq r \tag{36}$$

and

$$D\{P(x)(1 + x^{p+q})\} \geqq p + q \qquad \text{for all } 0 < p, \quad q \leqq r \tag{37}$$

where $Q_s(x)$ is the coset leader of the polynomial $\delta^s(x) + P(x)(1 + x^s)$.

3.1 *Scheme A — General Approach*

A decoder which corrects up to t errors while utilizing a code with minimum distance d_m can also detect up to $d = d_m - (t + 1)$ errors. Thus, if every error pattern generated by slippage $s \leqq r$ has weight between $t + 1$ and d, it will certainly be detected. It follows that these error patterns cannot be in $\mathfrak{C}$. Based on this sufficient condition one can design $P(x)$ accordingly. Suppose

$$d \geqq W[\delta^s(x) + P(x)(1 + x^s)] \geqq t + 1 \tag{38}$$

for all random polynomials $\delta^s(x)$ for $0 < s \leqq r$. Let

$$P(x) = \sum_{i=0}^{n-1} P_i x^i \tag{39}$$

and

$$T_s(x) = \sum_{i=0}^{n-s-1} (P_i + P_{i+s})x^{i+s}, \qquad 0 < s \leqq r. \tag{40}$$

Then

$$d - s \geqq W(T_s(x)) \geqq t + 1, \qquad 0 < s \leqq r, \tag{41}$$

is necessary and sufficient for (38) to hold. In particular, it is necessary that $W[T_s(x)] \geqq t + 1$. It follows that the best choice for $P(x)$, in the sense of maximizing r, is for

$$W[T_s(x)] = t + 1, \qquad \text{for all } 0 < s \leqq r. \tag{42}$$

Those $P(x)$ that satisfy (42) will be called optimal. In general, for certain special codes it is possible to find other schemes which work even if $d - r \geqq t + 1$ is not satisfied. This possibility will be discussed later.

It can be shown by direct substitution* in (41) that the following $P(x)$ are optimal.

$$P(x) = \sum_{\sigma=\sigma_0}^{[t/2]} x^{\sigma(r+1)-\sigma_0} + x^{n-1} \tag{43}$$

if

$$r + [t/2](r + 1) - \sigma_0 < n - 1$$

where

$$\sigma_0 = 1 + 2[t/2] - t \tag{44}$$

and $[t/2]$ represents the integer part of $t/2$.

TABLE III

Code	d_m	Number of Errors Corrected t	Levy's Result (Max. Number of Sync Loss Detected)	Optimal Result (Max. Number of Sync Loss Detected)
(23,12)	7	1	2	3
(127,85)	13	1	8	9
		2	5	7
		3	2	5
(255,191)	17	1	12	13
		2	9	11
		3	6	9
		4	3	7
(255,163)	25	1	20	21
		2	17	19
		3	14	17
		4	11	15
		5	8	13

Levy[9] has found a set of $P(x)$ for the purpose of detecting synchronization loss. Table III compares Levy's results† with some of the optimal results just derived. It is interesting to note that appreciable improvements are obtained by optimal $P(x)$.

It remains to show that the polynomials $P(x)$ of (43) possess sync-recovery capability. The following theorem characterizes the extent of slippage the code can correct provided that $d - r \geqq t + 1$.

* See Appendix.

† See Ref. 9, page 11, Table 1. Note $\delta \geqq t + 1$ is necessary to detect sync loss. Table III is constructed by letting $\delta = t + 1$.

Theorem 4: To every t error-correcting cyclic code there exists a coset code with sync-recovery capability of at least

$$r < \frac{n - k - t + [t/2]}{2 + [t/2]} \textit{ bits}$$

provided that:

(*i*) *the coset code is generated by*

$$P(x) = \sum_{\sigma=\sigma_0}^{[t/2]} x^{\sigma(r+1)-\sigma_0} + x^{n-1}$$

(*ii*) $d - r \geqq t + 1$

(*iii*) $r + [t/2](r + 1) - \sigma_0 < n - 1$

where

$$\sigma_0 = 1 + 2[t/2] - t.$$

Proof: Since $P(x)$ has sync-detection capability r when $d - r \geqq t + 1$, all one has to check is that the syndrome of error patterns due to bit loss are different from those due to bit gain.

From (37) one requires

$$\begin{aligned} D\{P(x)(1 + x^{p+q})\} &= D\left\{\left(\sum_{\sigma=\sigma_0}^{[t/2]} x^{\sigma(r+1)-\sigma_0} + x^{n-1}\right)(1 + x^{p+q})\right\} \\ &= D\{1 + x^{p+q} + \cdots + x^{[t/2](r+1)-\sigma_0+p+q} \\ &\qquad + x^{n+1}\} \leqq p + q, \qquad 1 \leqq p + q \leqq r. \end{aligned} \tag{45}$$

Let the generator polynomial be denoted

$$g(x) = \sum_{i=0}^{n-k} g_i x^i ;$$

then

$$x^{-1} g(x) = g_0 x^{-1} + \sum_{i=1}^{n-k} g_i x^{i-1} .$$

In general, one may assume $g_0 = g_{n-k} = 1$, and because $x^n = 1$, we have

$$x^{-1} g(x) = x^{n-1} + \sum_{i=1}^{n-k} g_i x^{i-1} = x^{n-1} + Q(x)$$

or

$$x^{n-1} = x^{n-1} g(x) + Q(x).$$

Hence

$$\{x^{n-1}\} = Q(x),$$

but as $Q(x)$ has a degree exactly $n - k - 1$; it follows that

$$D\{x^{n-1}\} = n - k - 1.$$

Therefore, (45) has a degree $n - k - 1$ if the next highest order term, $x^{[t/2](r+1)+p+q-\sigma_0}$ is always of a degree less than $n - k - 1$, for all $1 \leqq p$, $q \leqq r$; i.e.,

$$[t/2](r + 1) + p + q - \sigma_0 < n - k - 1, \quad 1 \leqq p,\ q \leqq r. \tag{46}$$

But, max $(p + q) = 2r$, i.e.,

$$[t/2](r + 1) + 2r - \sigma_0 < n - k - 1$$

or

$$r < \frac{n - k - 1 + \sigma_0 - [t/2]}{2 + [t/2]} .$$

Recall that $\sigma_0 = 1 + 2[t/2] - t$, and therefore

$$r < \frac{n - k - t + [t/2]}{2 + [t/2]}$$

is sufficient to satisfy (45).

Example: Find $P(x)$ for (15,6) BCH code where $d_m = 6$.

(i) If the code is used for single error correction, $t = 1$ and

$$d - r = d_m - (t + 1) - r \geqq t + 1.$$

Therefore, $r \leqq 2$, but

$$\frac{n - k - t + [t/2]}{2 + [t/2]} = (9 - 1)/2 = 4$$

so that by Theorem 4, $P(x) = 1 + x^{14}$ is a valid pattern for sync correction up to 2 bits.

(ii) If the code is used for double error correction $t = 2$ and $d = 3$. Thus, $d - r = 3 - r \geqq t + 1 = 3$. Therefore, $r = 0$, so that it is impossible to use this method but there are still other possibilities which will be discussed in the next section.

3.2 *Scheme B — Special Case*

The scheme just derived is applicable to all binary cyclic codes provided that $d_m > 2t + 2$, where t is the number of errors the decoder actually corrects, and that the scheme is not applicable to those systems in which the decoder is designed to correct $t = [(d_m - 1)/2]$ errors.

In this section, a different technique is developed. Instead of requiring that $d_m > 2t + 2$, this technique requires that a set of conditions on the code structure be satisfied. These conditions are almost always satisfied by *Bose-Chaudhuri-Hocquenghem* codes that correct more than two errors. For such codes it will be shown that it is always possible to obtain a coset code that has sync-recovery capability of at least one bit, even if the decoder is designed to correct errors up to the guaranteed error-correcting capability of the code, i.e., $t = [(d_m - 1)/2]$.

Definition 9: The weight of a coset is defined as the weight of its coset leader.

Definition 10: Code C_2 is said to be a descendant of code C_1 if

$$a \in C_1 \Rightarrow a \in C_2, \qquad C_1 \neq C_2$$

and is denoted by the notation

$$C_1 \subset C_2 .$$

Theorem 5: $C_1 \subset C_2$ if and only if $g_2(x) \mid g_1(x)$
where
$g_1(x)$ and $g_2(x)$ are the generators of codes C_1 and C_2, respectively.

Proof: (*i*) Since $g_2(x)$ itself is a code word of C_2, if $g_2(x)$ does not divide $g_1(x)$ then $g_1(x) \notin C_2 \Rightarrow C_1 \not\subset C_2$; a contradiction.

(*ii*) If $g_2(x) \mid g_1(x)$, then let

$$\begin{aligned} r(x) &= g_1(x)/g_2(x) \\ w(x) \in C_1 \Leftrightarrow w(x) &= g_1(x)f(x) \\ &= g_2(x)r(x)f(x) \\ &\Leftrightarrow w(x) \in C_2 . \end{aligned}$$

Hence, $C_1 \subset C_2$, by definition.

Theorem 6: Suppose code C_1 corrects t_1 errors and code C_2 corrects t_2 errors. If $C_1 \subset C_2$ and $t_1 > t_2$, then the coset Ω of C_1 that the code word $K(x) \in (C_2 - C_1)$ belongs, must have a weight of at least $2t_2 + 1$.

Proof: By definition $(C_2 - C_1)$ is nonempty, so there exists $K(x) \in C_2$ but $K(x) \notin C_1$. All the elements, of the coset Ω of C_1 to which $K(x)$ belongs, must be of the form

$$K(x) + w(x), \qquad w(x) \in C_1$$

but, $C_1 \subset C_2$, so

$$w(x) \in C_2 .$$

It follows that

$$K(x) + w(x) \in C_2 ;$$

thus, the weight of $K(x) + w(x)$ must be at least $2t_2 + 1$, as C_2 is a t_2-error-correcting code. It follows that Ω must have weight at least $2t_2 + 1$.

Theorem 7: Let Ω be a coset of a code C, $K(x)$ an element $\in \Omega$, and suppose the weight of Ω is R. Then the coset Ω_1 to which $K(x) + x^i$ belongs must have weight not less than $R - 1$.

Proof:

$$\begin{aligned}\Omega_1 &= \{(K(x) + x^i) + w_k(x) : w_k(x) \in C\}\\ &= \{(K(x) + w_k(x)) + x^i : w_k(x) \in C\}\\ &= \{\Omega(x) + x^i : \Omega(x) \in \Omega\}.\end{aligned}$$

Since Ω has weight R, each element of Ω_1 differs from an element in Ω by exactly one term. It follows that the weight of Ω_1 is at least $R - 1$.

The following several lemmas, in associated with Theorems 6 and 7, are essential to Theorem 8 which is the basis for Scheme B.

Lemma 2: $$\{K(x)\} \neq \{x^{n-1}K(x)\} \qquad [\text{mod } g_1(x)]$$

if

$$K(x) \in C_2 - C_1, \qquad C_1 \subset C_2$$

and

$$g_1(x)/g_2(x) = r(x) \nmid (1 + x)$$

where $g_1(x)$ and, $g_2(x)$ are generator polynomials of C_1, C_2, respectively.

Proof: By hypothesis,

$$(1 + x) \nmid r(x). \tag{47}$$

Assume

$$\{K(x)\} = \{x^{n-1}K(x)\} \qquad [\text{mod } g_1(x)]$$

i.e.,

$$\{(1 + x)K(x)\} = 0 \qquad [\text{mod } g_1(x)] \tag{48}$$

or

$$(1 + x)K(x) = f(x)g_1(x).$$

Now,

$$(1 + x) \mid f(x) \Rightarrow K(x) = f_1(x)g_1(x)$$
$$\Rightarrow K(x) \in C_1 \text{ ; a contradiction.}$$

It follows that $(1 + x) \nmid f(x)$. (49)

Since $K(x) \in C_2$,

i.e.,

$$K(x) = f_2(x)g_2(x),$$

by the assumption of (48)

$$\begin{aligned}(1 + x)K(x) = (1 + x)f_2(x)g_2(x) &= f(x)g_1(x)\\ &= f(x)g_2(x)r(x)\end{aligned}$$

or

$$(1 + x)f_2(x) = f(x)r(x). \tag{50}$$

In view of (47) and (49), and by the unique factorization theorem, (50) cannot be satisfied. The lemma follows by contradiction.

Lemma 3: $\{K(x) + 1\} \neq \{x^{n-1}K(x) + x^{n-1}\}$ $[\mathrm{mod}\ g_1(x)]$ *if*

$$K(x) \in C_2 - C_1 \quad \textit{and} \quad C_1 \subset C_2 .$$

Proof: Assume the contrary.
Then

$$\begin{aligned}K(x) + 1 + x^{n-1}K(x) + x^{n-1} &= f(x)g_1(x)\\ &= f(x)g_2(x)r(x)\end{aligned} \tag{51}$$

but

$$K(x) \in C_2 \Rightarrow g_2(x) \mid K(x).$$

The right-hand side is divisible by $g_2(x)$ but not the left-hand side unless

$$g_2(x) \mid (1 + x^{n-1}).$$

The code generated by $g_2(x)$ has natural length n so that $g_2(x) \mid (1 + x^n)$ but not any $1 + x^k$, $k < n$. Therefore, (51) cannot hold and the lemma follows.

Lemma 4: $\{K(x)\} \neq \{x^{n-1}K(x) + x^{n-1}\} \quad [\text{mod } g_1(x)]$
if

$$K(x) \in C_2 - C_1, \qquad C_1 \subset C_2 .$$

Proof: Assume the contrary.
Then

$$x^{n-1}K(x) + K(x) + x^{n-1} = f(x)g_2(x)r(x).$$

The left-hand side is not divisible by $g_2(x)$, since $g_2(x) \mid K(x)$, so the lemma follows.

Lemma 5: $\{K(x) + 1\} \neq \{x^{n-1}K(x)\} \quad [\text{mod } g_1(x)]$
if

$$K(x) \in C_2 - C_1, \qquad C_1 \subset C_2 .$$

Proof: Similar to Lemma 4.

Theorem 8: Suppose code C_1 and C_2, with generator $g_1(x)$ and $g_2(x)$, correct t_1 and t_2 errors, respectively, and $2t_2 > t_1$. If $C_1 \subset C_2$,

$$g_2(x)/g_1(x) = r(x) \nmid (1 + x)$$

and if $K(x) \in C_2 - C_1$, then the pattern

$$P(x) = \left\{\frac{K(x)}{1 + x}\right\},$$

if $K(x)$ has even weight, or

$$P(x) = \left\{\frac{K(x) + 1}{1 + x}\right\},$$

if $K(x)$ has odd weight, defines a coset code of C_1 with sync-recovery capability of at least one bit.

Proof: First note that such $P(x)$ always exists. Now with the $P(x)$ used to define the coset code, the error pattern for one-bit loss in sync is

$$\begin{aligned} &\{P(x)(1 + x) + \delta^1(x)\} \\ &\qquad = \{K(x) + \delta^1(x)\} && \text{if } K(x) \text{ is even} \\ &\qquad = \{K(x) + 1 + \delta^1(x)\} = \{K(x) + \delta^1(x)\} && \text{if } K(x) \text{ is odd.} \end{aligned}$$

That is, the syndromes are either $\{K(x)\}$ or $\{K(x) + 1\}$. By Theorem 6, $\{K(x)\}$ does not belong to a correctable coset of C_1 since $2t_2 > t_1$. The coset corresponding to $\{K(x) + 1\}$ has weight equal to or greater than $2t_2$, by Theorem 7, so that it is not correctable.

Similarly, the error pattern for one-bit gain is

$$\begin{aligned}\{P(x)(1+x^{n-1})&+x^{n-1}\delta^1(x)\}\\ &=\{x^{n-1}[P(x)(1+x)+\delta^1(x)]\}\\ &=\{x^{n-1}K(x)+x^{n-1}\delta^1(x)\} && \text{if } K(x) \text{ is even}\\ &=\{x^{n-1}(K(x)+1)+x^{n-1}\delta^1(x)\}\\ &=\{x^{n-1}K(x)+x^{n-1}\delta^1(x)\} && \text{if } K(x) \text{ is odd.}\end{aligned}$$

That is, the syndromes are either

$$\{x^{n-1}K(x)+x^{n-1}\} \qquad \text{or} \qquad \{x^{n-1}K(x)\}$$

By Theorem 6 and Lemma 1, $\{K'(x)\} = \{x^{n-1}K(x)\}$ is not in any of the correctable cosets of C_1, and by Theorem 7 and Lemma 1,

$$\{x^{n-1}K(x)+x^{n-1}\} = \{x^{n-1}(K(x)+1)\}$$

is not correctable either. It follows that $P(x)$ satisfies the first condition that all the error patterns of one-bit slippage generated by $P(x)$ can not be in any of the correctable cosets of C_1 so that C_1 has sync-detection capability of at least one bit.

By Lemmas 2, 3, 4, and 5 the cosets corresponding to bit-loss patterns can not be the same as the bit-gain patterns. Thus the second condition is satisfied. It follows that code C_1 has sync-recovery capability of at least one bit.

The search for $K(x)$ is very simple since $g_2(x) \in C_2$, $g_2(x) \notin C_1$ so that we may use $g_2(x)$ for $K(x)$ in every instance.

Thus the procedure to find a coset code for use with Scheme B is as follows.

(*i*) Find a code, C_2, of the given code C_1 such that $g_2(x)/g_1(x) = r(x) \nmid (1+x)$ and that $2t_2 > t_1$.

(*ii*) Use

$$\left\{\frac{g_2(x)}{1+x}\right\} = P(x)$$

to generate the desired coset code if the weight of $g_2(x)$ is even and use

$$\left\{\frac{g_2(x)+1}{1+x}\right\} = P(x)$$

if $g_2(x)$ has odd weight.

From Theorem 8, it is easy to see that this procedure applies to all Bose-Chaudhuri-Hocquenghem codes whenever $2t_2 > t_1$, and many other algebraic codes.

The following example shows how to apply Theorem 8 to Bose-Chaudhuri-Hocquenghem codes.

The polynomial and the associated sync-loss error syndromes $P(x)$ for a (15,5) BCH triple-error-correcting code generated by

$$g_1(x) = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1)$$

can be found as follows:

If α is a root of $x^4 + x + 1$ then α^5 is a root of $x^2 + x + 1$ (see tables of Marsh[13] or Peterson[12]) so that

$$r(x) = (x^2 + x + 1) \nmid (1 + x)$$

$$g_2(x) = \frac{g_1(x)}{r(x)} = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)$$

generates a double error-correcting BCH code.

Observe that

$$t_1 = 3, \qquad t_2 = 2$$

therefore,

$$2t_2 > t_1 ,$$

hence, all the requirements for Scheme B are satisfied. Set

$$K(x) = g_2(x),$$

and since $g_2(x)$ has odd weight use

$$P(x) = \left\{\frac{K(x) + 1}{1 + x}\right\} = \left\{\frac{x^8 + x^7 + x^6 + x^4}{1 + x}\right\}$$
$$= x^7 + x^5 + x^4 .$$

The syndromes are:

(*i*) bit loss

$$\{K(x) + 1\} = x^8 + x^7 + x^6 + x^4$$

or

$$\{K(x)\} = x^8 + x^7 + x^6 + x^4 + 1.$$

(*ii*) bit gain

$$\{x^{n-1}K(x)\} = x^9 + x^6 + x^5 + x^4 + x + 1$$

or

$$\{x^{n-1}K(x) + x^{n-1}\} = x^7 + x^6 + x^5 + x^3.$$

It can be verified that all the syndromes listed do not belong to correctable cosets of the given code, and the syndromes are obviously all different, so the modified (15,5) code has sync-correction capability of one bit.

Notice that for this code

$$d_m = 7, \qquad t_1 = 3;$$

therefore,

$$d = d_m - (t + 1) = 7 - 4 = 3$$

and

$$t_1 + 1 = 4.$$

The condition, $d - r \geqq t + 1$, is not satisfied so that Scheme A is not applicable.

3.3 *Implementation*

By the Euclidean division algorithm, one writes

$$(1 + x^s)P(x) = g(x)F(x) + R_s(x)$$

where $D[R_s(x)] < D[g(x)] = n - k$.

From (28), the s-bit-loss syndromes are

$$\{\delta^s(x) + (1 + x^s)P(x)\} = \{\delta^s(x)\} + \{(1 + x^s)P(x) = \delta^s(x) + Rs(x)$$

(Since $D[\delta^s(x)] < D[g(x)]$) for all $0 < s \leqq r$.

It follows that all possible s-bit-loss syndromes have the same high order $n - k - s$ terms [namely, the high order $n - k - s$ terms of $Rs(x)$] and the remaining low order s-terms assume all possible 2^s combinations. Thus, the 2^s possible syndromes for an s-bit-loss can be detected by a single *and* gate that recognizes the high order $n - k - s$ terms of $R_s(x)$.

According to (28) and (30), s-bit-gain syndromes are the same as the s-bit-loss syndromes multiplied by x^{n-s}. It follows that a bit-loss recognition device, as mentioned above, can also be used to test bit-gain syndromes. This can be done by transforming bit-gain syndromes to bit-loss syndromes through multiplication by x^s, $0 < s \leqq r$, then testing the resultant syndromes with the bit-loss recognition device. Such a device takes r *and* gates and is applicable to both Scheme A and B.

IV. CYCLIC CODES FOR DETECTION ONLY

Some parts of the subject discussed in this section have been investigated in the literature.[8] Here, a different point of view is presented.

If a cyclic code is used for error detection only, then any error pattern due to sync loss or gain can be detected so long as the erroneous words generated by sync loss are not in the code space. If r is the maximum amount of slippage possible (such that the misframed words are not code words), the code is usually said to possess comma-free freedom r. Codes having the property that $| r | = [(n + 1)/2]$, are called "comma free".

Consider a coset code C generated by $P(x)$ and designed to detect sync loss or gain. To assure that such error patterns are detectable, one requires, by (28) and lemma 1,

$$\{ (1 + x^{\beta})P(x) + \delta^{\beta}(x)\} \neq 0 \tag{52}$$

since $\mathcal{C} = 0$ because every coset is not correctable for error-detecting codes.

As any syndrome has a degree which is at most $n - k - 1$, and $\delta^{\beta}(x)$ is an arbitrary polynomial of degree $\beta - 1$, it is always possible, for any $P(x)$ to have

$$\{ (1 + x^{\beta})P(x) + \delta^{\beta}(x)\} = 0 \qquad \text{if} \qquad \beta > n - k - 1.$$

It follows that the comma-free freedom can never exceed $n - k - 1$. Hence, we have the following theorem.

Theorem 9: The comma-free freedom of any cyclic code cannot exceed $n - k - 1$.

One sees that by letting $P(x) = 1$; (52) now reads

$$\{ (1 + x^{\beta})\cdot 1 + \delta^{\beta}(x)\} = \{1 + x^{\beta} + \delta^{\beta}(x)\} = x^{\beta} + \delta^{\beta}(x). \tag{53}$$

The term x^{β} cannot be canceled by $\delta^{\beta}(x)$ provided $\beta \leqq n - k - 1$. It follows that the upper bound on comma-free freedom described by Theorem 9 can be met. Hence, we have the following result.

Theorem 10: The comma-free freedom of any cyclic code can be made as large as $n - k - 1$.

According to Theorem 10, the comma-free freedom r cannot be greater than $n - k - 1$. Thus, if $k \geqq (n - 1)/2$, it is impossible to detect all the sync loss for $n - k \leqq r \leqq k$; on the other hand, if $k < (n - 1)/2$ then the interval between $n - k$, k does not exist. Hence, every slippage can be detected. It follows that

Corollary 1: An (n,k) cyclic code can be made comma-free if and only if $k < (n - 1)/2$.

The above results have been proved in a different manner by J. J. Stiffler.[8]

Clearly, for strictly error-detecting codes, if a received message is not a code word, no distinction can be made to decide whether the error is caused by additive noise or is due to sync loss. However, statistical decisions still can be made accurately by observing the number and frequency of word errors, and, if it is concluded that a sync loss has occurred, sync can be recovered by sliding the word frame until the number of errors observed abruptly reduces to a predetermined level. The penalty for such a process is, of course, the time delay and the loss of data.

V. CONCLUSIONS

Techniques for automatic word synchronization recovery are presented. The techniques are useful if the slippage of word framing is not large, which is presumably the usual case.

An upper bound on the synchronization recovery capability for any cyclic code is found. It is shown that, for recovery to be possible, the amount of slippage in bits, r, cannot exceed $(n - k - 1)/2$. It is also shown that the synchronization-loss detection capability of any cyclic code is upper-bounded by $n - k - 1$ bits and furthermore, the bound can be met.

For shortened cyclic codes, the technique has five valuable features:

(*i*) No additional redundancy is required if the code corrects more than two errors, and only two additional check bits are required if the code corrects one or two errors. The generation of such check bits is simple.

(*ii*) The normal error-correcting ability of the code is not reduced when synchronization is maintained.

(*iii*) The correction of synchronization loss can be accomplished even in the presence of additive noise.

(*iv*) The time delay required before proper framing is restored is small, usually one-word time.

(*v*) The implementation is very simple.

The technique has been successfully applied to an existing error control unit[14] which utilizes a triple-error-correcting (200,175) code. The net cost of the additional hardware is about 20 transistors. The circuit corrects at least one-bit synchronization loss and, with diminishing probabilities, corrects larger sync losses as well.

For cyclic codes which are not shortened, necessary and sufficient conditions for the existence of a suitable coset code without additional redundancy for the recovery of synchronization loss are derived; and a class of optimal codes is given.

Two schemes are presented for finding such coset codes. The first scheme, called Scheme A, applies to any cyclic code whose minimum distance is greater than $2t + 2$, where t is the number of random errors the decoder actually corrects.

Such a scheme is usually applicable to data systems with a reverse channel in which case high error-detecting ability is utilized to obtain very-high-accuracy transmission. This requires that the error-correcting ability of the code be reduced in favor of the detecting ability. That is to say, in such a case, the minimum distance of the code is often greater than $2t + 2$.

For systems using forward-acting error correction only, the error-correcting ability of the code is usually exploited fully so that the requirements of Scheme A may not be met. A special technique, called Scheme B, are developed for such situation. Instead of requiring $d_m > 2t + 2$, Scheme B requires a set of conditions which are almost always satisfied by Bose-Chaudhuri-Hocquenghem codes that correct more than two errors. Both schemes have the advantages mentioned earlier for shortened cyclic codes except there is no noise tolerance.

Word synchronization loss is a catastrophic failure in error-control systems. The techniques herein described offer a solution to this problem for all binary cyclic codes with negligible cost in hardware, in time delay, and without loss in transmitting efficiency in many cases.

VI. ACKNOWLEDGMENTS

The author would like to express his gratitude to his thesis advisors, Professor E. J. McCluskey, Jr. and Professor A. McKellar, for their encouragement and advice. The author is also deeply indebted to H. O. Burton for many stimulating discussions and for assisting with the preparation of this manuscript.

Thanks are due to E. J. Weldon, Jr. for many references he provided and to P. G. Neumann for his valuable suggestions.

APPENDIX

The Verification that $P(x)$ is Optimal

$$P(x) = \sum_{i=0}^{n-1} P_i x^i = \sum_{\sigma=\sigma_0}^{[t/2]} x^{\sigma(r+1)-\sigma_0} + x^{n-1} \tag{54}$$

where

$$r + [t/2](r + 1) - \sigma_0 < n - 1 \tag{55}$$

and

$$\sigma_0 = 1 + 2[t/2] - t.$$

Now

$$T_s(x) = \sum_{i=0}^{n-s-1} (P_i + P_{i+s})x^{i+s} \qquad 0 < s \leqq r. \tag{56}$$

Consider

$$\begin{aligned}(1 + x^s)P(x) &= \left(\sum_{\sigma=\sigma_0}^{[t/2]} x^{\sigma(r+1)-\sigma_0} + x^{n-1}\right)(1 + x^s) \\ &= \sum_{\sigma=\sigma_0}^{[t/2]} (x^{\sigma(r+1)-\sigma_0} + x^{\sigma(r+1)-\sigma_0+s}) \\ &\qquad + x^{n-1} + x^{s-1}.\end{aligned} \tag{57}$$

Note that:

(*i*) $\sigma(r + 1) - \sigma_0 \neq \sigma(r + 1) - \sigma_0 + s$ for all $0 < s \leqq r$.

(*ii*) The exponent of the highest order term under the summation sign, $[t/2](r + 1) - \sigma_0 + s < [t/2](r + 1) - \sigma_0 + r < n - 1$ by (55).

Therefore, no terms of (57) will be canceled for all $0 < s \leqq r$.

Note $T_s(x)$ is the same as the polynomial (57) with the lower order terms $x^0, x^1, \cdots, x^{s-1}$ removed so that the number of nonzero terms (hence the weight) of $T_s(x)$ is equal to the total number of terms of (57) minus the number of nonzero terms with exponent $s - 1$ or less.

Case 1: t is odd.

$$\sigma_0 = 1 + 2[t/2] - t = 1 + 2\frac{t-1}{2} - t = 0.$$

There are two nonzero terms in (57) which have exponents no greater than $s - 1$, namely $x^{\sigma_0(r+1)-\sigma_0} = x^0$ and x^{s-1}. Therefore, the weight of $T_s(x)$

$$\begin{aligned}&= 2([t/2] + 2) - 2 = 2\left(\frac{t-1}{2} + 2\right) - 2 \\ &= t + 1.\end{aligned}$$

Case 2: t is even.

$$\sigma_0 = 1 + 2[t/2] - t = 1 + 2(t/2) - t = 1.$$

There is only one nonzero term, namely, x^{s-1} that has an exponent no greater than $s - 1$. Therefore, the weight of $T_s(x)$

$$= 2([t/2] + 1) - 1 = 2(t/2 + 1) - 1 = t + 1.$$

REFERENCES

1. Gilbert, E. N., Synchronization of Binary Messages, IEEE Trans. Inform. Theor., *IT-6,* No. 4, September, 1960, pp. 470–477.
2. Neumann, P. G., On a Class of Efficient Error Limiting Variable Length Codes, IEEE Trans. Inform. Theor., *IT-8,* No. 5, September, 1962, pp. 260–260.
3. Neumann, P. G., Efficient Error Limiting Variable Length Codes, *Ibid, IT-8,* No. 4, July, 1962, pp. 292–304.
4. Neumann, P. G., Error Limiting Coding Using Information Loseless Sequential Machine, *Ibid, IT-10,* No. 2, April, 1964, pp. 108–115.
5. Stiffler, J. J., Synchronization of Telemetry Codes, IEEE Trans. Space Electron. and Telemet., *SET-8,* No. 3, June, 1962, pp. 112–117.
6. Barker, R. H., Group Synchronizing of Binary Digital Systems, *Communication Theory,* W. Jackson, Ed., Academic Press, 1953, pp. 273–287.
7. Sellers, F. F., Bit Loss and Gain Correction Code, IEEE Trans. Inform. Theor., *IT-8,* No. 1, January, 1962, pp. 35–38.
8. Stiffler, J. J., Comma-Free Error-Correcting Codes, IEEE Trans. Inform. Theor., *IT-11,* No. 1, January, 1965, pp. 107–112.
9. Levy, J. E., Self-Synchronizing Codes Derived from Binary Cyclic Codes, Unpublished Report, ADCOM, Inc., Cambridge, Mass., March, 1965.
10. Bose, R. C. and Ray-Chaudhuri, D. K., On a Class of Error-Correcting Binary Group Codes, Information and Control, No. 3, 1960, pp. 68–79.
11. Hocquenghem, A., Codes Correcteurs d'erreurs, Chiffres 2, 1959, pp. 147–156.
12. Peterson, W. W., *Error-Correcting Codes,* MIT Press, 1961.
13. Marsh, R. W., *Table of Irreducible Polynomials Over GF(2) Through Degree 19,* NSA, Washington, D.C., 1957.
14. Burton, H. O., and Weldon, E. J., An Error Control System for Use With a High-Speed Voiceband Data Set, 1st IEEE Communication Conference Record, June, 1965, 485–488.
15. Frey, Jr., A. H., Message Framing and Error Control, IEEE Trans. Mil. Electron., *MIL-9,* No. 2, April, 1965, pp. 143–147.
16. Golomb, S. W., et al., Comma-Free Codes, Canada J. Math., *10,* No. 2, 1958.
17. Gorenstein, D., Peterson, W. W., and Zierler, N., Two Error Correcting Bose-Chaudhuri Codes are Quasi-Perfect, Information and Control, *3,* 1960, pp. 291–294.
18. Ullman, J. D., Near Optimal Single Synchronization Error-Correcting Codes, Technical Report No. 45, Digital System Labs, Department of E.E., Princeton University, March, 1965.

Increasing the Memory Capacity of the Digital Light Deflector by "Color Coding"

By J. G. SKINNER

(Manuscript received December 23, 1965)

Considerations are given to the use of a multiwavelength light source in the digital light deflector for the purpose of increasing the memory capacity of the system. An increase of a factor of twenty or more is theoretically possible when potassium tantalate niobate (KTN) is used for the optical switches in the digital light deflector. Practical considerations may, however, limit the increase to about a factor of ten.

I. INTRODUCTION

This paper describes a scheme for increasing the memory capacity of an optical beam deflecting device, such as a digital light deflector, by utilizing the full wavelength bandwidth of the system. For convenience the scheme is referred to as "color coding".

The digital light deflector[1] is designed to deflect a monochromatic light beam to any one of an array of positions on a memory plane. The absence or presence of an obstruction at the memory plane produces a yes or no signal in a detector placed behind the memory plane. The capacity of the system is determined by the ratio of the maximum angular deflection of the beam to the smallest angular deflection that can be resolved by the system. Increasing the capacity by increasing the maximum beam deflection requires lenses that can operate over a larger angular field of view. On the other hand, increasing the capacity by reducing the minimum resolvable angle requires larger diameter components. Assuming the diameter of the components and the angular field of view of the lenses are at their maximum values, then other means of increasing the capacity must be investigated.

The capacity may be increased by paralleling information through the system so that each resolvable angular location in space becomes a word. This may be accomplished by splitting the deflected monochromatic light beam into M channels, each with its own detector. However, due to imperfections in certain components in the digital light deflector, a small portion of the incident light beam is deflected to several incorrect locations on the memory plane and also appears as a general background illumination on the memory. This extraneous light is undesirable because it reduces the signal-to-noise ratio of the correctly deflected spot of light. A method of reducing the unwanted light is to place a mirror behind the memory plane to reflect the light back through the deflector system to a detector located in a position conjugate to the light source. It can be shown[2] that this system reduces the extraneous light. In order to parallel information through this type of system it is necessary to code the word in the memory system, reunite the word for transmission back through the light deflector and then separate again for detection. This can be accomplished with a monochromatic light beam by coding a single pulse into a sequence of pulses by placing the different memory planes at different distances from the exit of the digital light deflector. This scheme,[2] however, requires large aperture, high-precision lenses which are expensive.

Another possibility is to transmit simultaneously several monochromatic beams of slightly different frequencies so that the beam can be reunited after coding and still be capable of yielding the coded word. The presence or absence of a reflector for a given frequency at a given location yields a yes or no bit of information. This is shown schematically in Fig. 1. The increase in capacity attainable with color coding is the ratio of the bandwidth allowed by the system to the bandwidth required per channel.

There are several ways of constructing a color-coding system. One scheme is to place a number of narrowband reflection filters at the memory plane as shown in Fig. 2(a). However, although narrowband

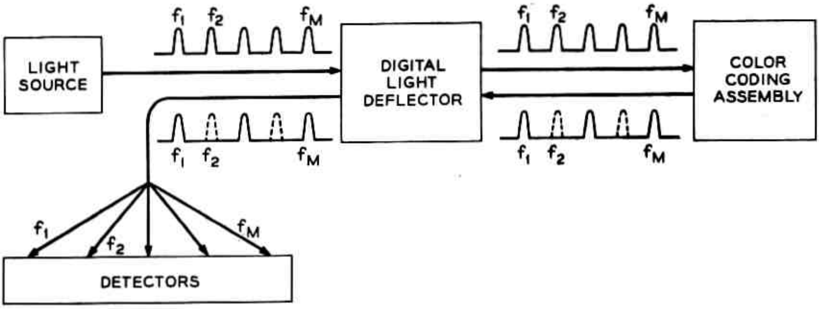


Fig. 1 — Schematic layout of the color-coding system.

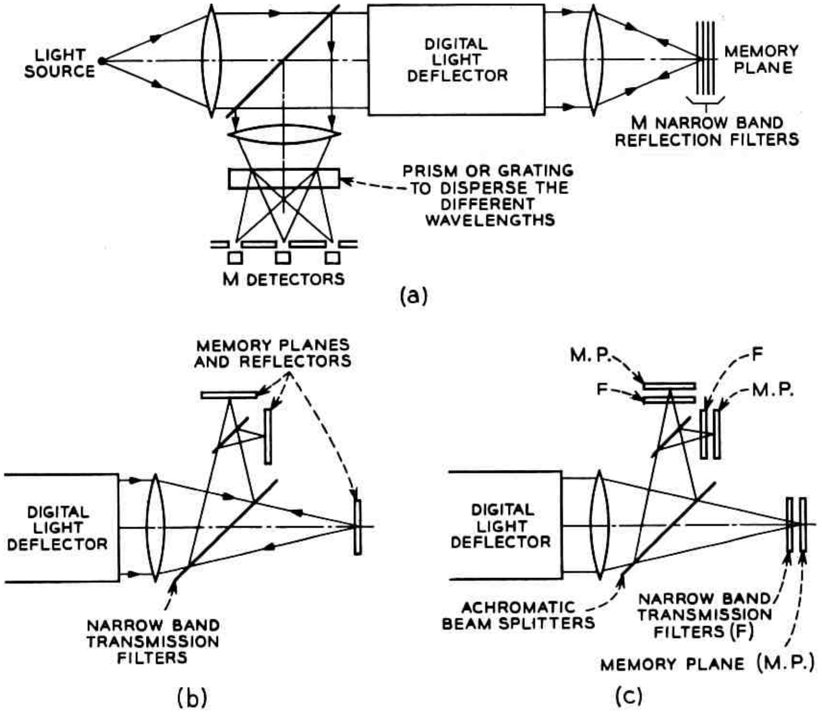

LIGHT SOURCE
DIGITAL LIGHT DEFLECTOR
MEMORY PLANE
M NARROW BAND REFLECTION FILTERS
PRISM OR GRATING TO DISPERSE THE DIFFERENT WAVELENGTHS
M DETECTORS
(a)
MEMORY PLANES AND REFLECTORS
DIGITAL LIGHT DEFLECTOR
NARROW BAND TRANSMISSION FILTERS
(b)
M. P.
F
F
M.P.
DIGITAL LIGHT DEFLECTOR
ACHROMATIC BEAM SPLITTERS
NARROW BAND TRANSMISSION FILTERS (F)
MEMORY PLANE (M.P.)
(c)

Fig. 2 — Different color-coding schemes.

reflection filters are theoretically possible, they are not presently available. Another scheme is to use narrow-band transmission filters which will transmit one channel and reflect the other channels to their appropriate locations. This is shown schematically in Fig. 2(b). This has the advantage that no energy is lost in the beam-splitting assembly but has the disadvantage that a filter requires a larger bandwidth for a given angular beam deflection when placed at an angle to the beam. In order to keep the bandwidth per channel to a minimum, it is necessary to use the filters normal to the beam. This arrangement requires an achromatic beam splitter to divide the beam into M channels. This is shown schematically in Fig. 2(c). The disadvantage of this scheme is that it reduces the beam intensity of each channel by a factor of M^2.

The calculations presented below are for narrow-band filters placed normal to the beam. If transmission filters were used, it would be necessary to tilt the filter at a small angle to avoid the unwanted light being reflected back to the detector. This small angular displacement of the filter has been ignored in these calculations. The calculations are based on the present design of the digital light deflector that employs potassium tantalate niobate (KTN) as the electro-optic material[3] for the light switches.[4]

II. BANDWIDTH OF THE DIGITAL LIGHT DEFLECTOR

The total bandwidth of the system is limited by the electro-optic elements that are used to rotate the plane of polarization of the light through an angle of 90°. Any error in this rotation allows light to leak through to the wrong locations in the memory plane. The light leakage is related to the phase difference $\Delta\Phi$ between the ordinary and the extraordinary rays at the exit face of the optical switch. Fig. 3 shows the parameters of the light beam relative to the optical switch. The value γ is the angle between the beam and the axis of the optical system, and α is the angle the intersection of the plane of incidence of the beam and the x,y plane makes relative to the x axis. The value of $\Delta\Phi$ to the second order in γ is given by[2,4]

$$\Delta\Phi = \left[1 + \frac{\left(\cos^2\alpha - \frac{1}{2}\right)\gamma^2}{n^2}\right]\frac{n^3}{\lambda}\pi l(g_{11} - g_{12})P^2 \qquad (1)$$

where n is the ordinary index of refraction, λ is the free space wavelength, l is the length of the optical switch, g_{11} and g_{12} are the electro-optic coefficients, and P is the lattice polarization of the KTN crystal.

With respect to the angular orientation α, the maximum change in $\Delta\Phi$ occurs at $\alpha = 0°$. Under this condition (1) reduces to

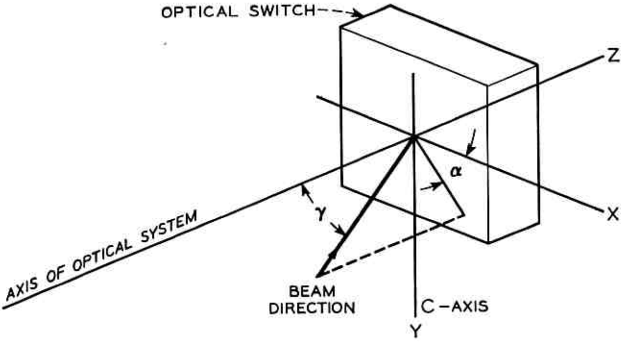


Fig. 3 — Coordinate axes showing the relation of the optical switch to the incident beam.

$$\Delta\Phi = \left[1 + \frac{\gamma^2}{2n^2}\right] \frac{\pi l n^3}{\lambda} (g_{11} - g_{12}) P^2. \tag{2}$$

We will assume that the optical switches are normally biased to $\Delta\Phi = N\pi$ and are switched to $\Delta\Phi_o = (N + 1)\pi$ for $\lambda = \lambda_0$ and $\gamma = 0°$. The error in $\Delta\Phi_o$, due to λ and γ being different from λ_0 and 0°, respectively, is

$$\delta\Phi = \frac{\pi l n_0^3}{\lambda_0} (g_{11} - g_{12}) P^2 \left[1 - \frac{\lambda_0 n^3}{n_0^3 \lambda}\left(1 + \frac{\gamma^2}{2n^2}\right)\right] \tag{3}$$

$$\delta\Phi = (N + 1)\pi\left[1 - \frac{\lambda_0 n^3}{\lambda n_0^3}\left(1 + \frac{\gamma^2}{2n^2}\right)\right] \tag{4}$$

where n_0 and n are the refractive indices at wavelengths λ_0 and λ, respectively, and are given by the equation[4]

$$n^2 = 1 + \frac{3.7994}{1 - (\lambda_s/\lambda)^2} \tag{5}$$

where $\lambda_s = 0.2012\ \mu$. The light leakage δ, due to the error in $\Delta\Phi_o$, is given by

$$\delta = I/I_0 = \sin^2 (\delta\Phi/2). \tag{6}$$

Combining (4), (5), and (6) we can obtain a plot of the allowed bandwidth, $\Delta\lambda = \lambda - \lambda_0$, for a given dc bias of the optical switch, expressed by N, and an allowed light leakage δ. The bandwidth has been plotted in Fig. 4 as a function of N for values of $\delta = 0.1$, 0.01, and 0.001, and a beam angle γ of 0.07 radians (i.e., 4°). This value of γ is a typical value that may be used in the digital light deflector; however, the term $\gamma^2/2n^2$ is negligible except for $\lambda = \lambda_0$.

A typical bias point for KTN optical switches is $N = 20$ and an acceptable light leakage in the digital light deflector, due to color coding, is $\delta = 0.01$. From Fig. 4 we see that the available bandwidth under these conditions is only 20 Å. The maximum bandwidth is at zero bias (i.e., $N = 0$) and equals 436 Å for a light leakage value of $\delta = 0.01$.

III. REQUIRED BANDWIDTH PER CHANNEL

The bandwidth required per channel depends on the type of filter used to separate the channels. We will consider a system using the Fabry-Perot type of narrow-band transmission filter[5] that consists of two reflecting surfaces separated by a suitable dielectric material. The

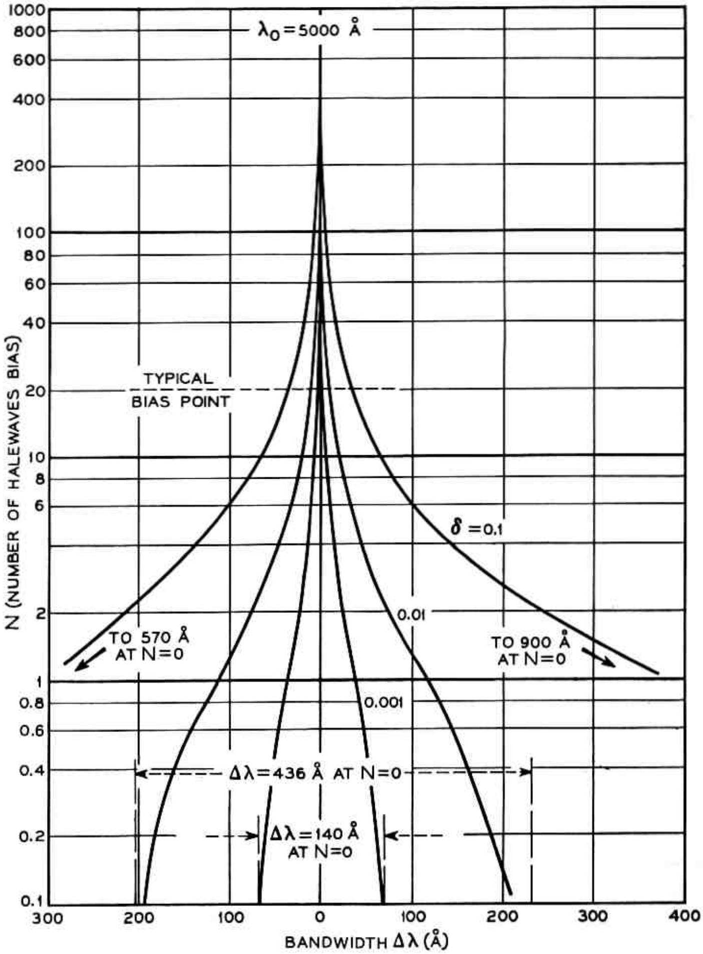


Fig. 4 — Frequency bandwidth of the KTN optical switches.

per cent transmission through such a filter depends on the wavelength and the angle of incidence of the beam. Since the beam will be deflected through a range of angles it is necessary to specify the minimum permissible transmission T, relative to the peak transmission, for each channel. It is also necessary to specify the maximum transmission T' of adjacent channels. The light reaching the detector from the memory plane passes through the filter twice, so that the single pass transmission values will be the square root of T and T'. The various parameters of the transmission curve are shown in Fig. 5; the values marked with a prime are parameters of the adjacent channel.

The peaks of the transmission for a Fabry-Perot filter occur at

$$\Omega = \frac{4\pi\mu h \cos\theta}{\lambda} = 2m\pi \qquad (m = 1,2,3, \cdots), \tag{7}$$

where μ is the refractive index of the dielectric between the two reflectors, h is the thickness of the dielectric, θ is the angle the beam travels through the dielectric relative to the normal of the surfaces of the filter, and λ is the free space wavelength of the beam.

The shape of the transmission curve is given by

$$\frac{I}{I_0} = \frac{1}{1 + F \sin^2 (\Omega/2)}$$

where $F = 4R/(1 - R)^2$ and R is the power reflection coefficient of the reflecting surfaces. If F is sufficiently large (i.e., high-reflectivity mirrors), then $\sqrt{F} = 4/\varepsilon$, where ε is the half bandwidth of the transmission curve.

We require that $I(T)/I_0 = \sqrt{T}$ at $\Omega = 2m\pi \pm \zeta$; therefore,

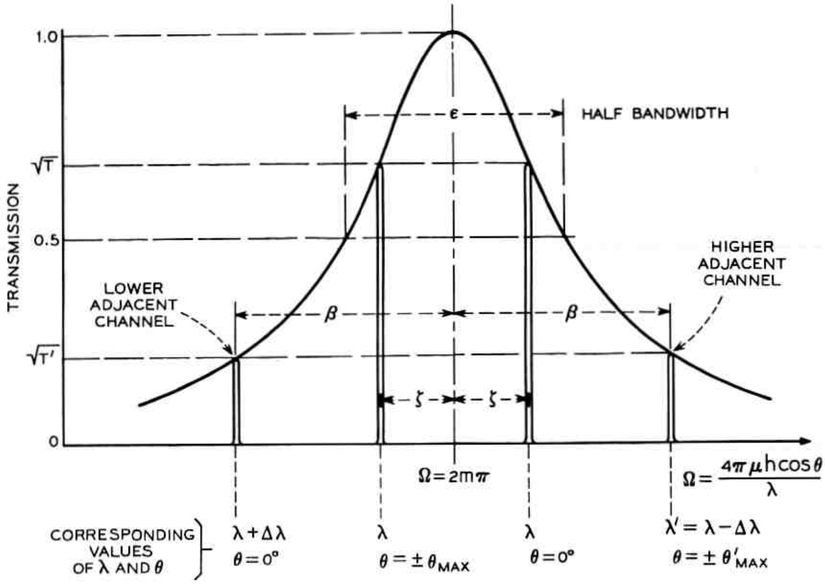


Fig. 5 — Transmission curve showing the various parameters involved.

$$\frac{I(T)}{I_0} = \sqrt{T} = \frac{1}{1 + F \sin^2\left(\frac{2m\pi + \zeta}{2}\right)} = \frac{1}{1 + F\sin^2(\zeta/2)}. \qquad (9)$$

Approximating $\sin(\zeta/2)$ with $(\zeta/2)$ for small values of ζ, we obtain

$$2\zeta = \frac{4}{\sqrt{F}} \sqrt{\frac{1}{\sqrt{T}} - 1} = \varepsilon t \qquad (10)$$

where

$$t = \sqrt{\frac{1}{\sqrt{T}} - 1}\,.$$

Similarly, the requirement that $I(T')/I_0 = \sqrt{T}$ at $\Omega = 2m\pi \pm \beta$ yields

$$2\beta = \varepsilon t' = 2\zeta t'/t \qquad (11)$$

where

$$t' = \sqrt{\frac{1}{\sqrt{T'}} - 1}\,.$$

From (7) and Fig. 5,

$$2m\pi - \zeta = 4\pi\mu h \cos \theta_{\max}/\lambda \qquad (12)$$

$$2m\pi + \beta = 4\pi\mu h \cos \theta_{\max}'/\lambda' \qquad (13)$$

$$2m\pi + \zeta = 4\pi\mu h/\lambda. \qquad (14)$$

From (12) and (14),

$$2\zeta = (2m\pi - \zeta)(1 - \cos \theta_{\max})/\cos \theta_{\max}\,. \qquad (15)$$

From (12) and (13), and noting that $\theta_{\max} = \theta_{\max}'$,

$$\zeta + \beta = (2m\pi - \zeta)\,\lambda \left(\frac{1}{\lambda'} - \frac{1}{\lambda}\right). \qquad (16)$$

Combining (11), (15), and (16), and expressing the bandwidth required between channels as $(\Delta\lambda) = \lambda - \lambda'$, we obtain

$$(\Delta\lambda) = \lambda(1 - \cos \theta_{\max})(t + t')/2t \cos \theta_{\max}\,. \qquad (17)$$

For small angles of $\theta_{\max}$, (17) may be written as

$$(\Delta\lambda) = \lambda\theta_{\max}^2(t + t')/4t. \qquad (18)$$

This equation shows that the required bandwidth per channel increases rapidly as the maximum beam deflection angle is increased.

The half bandwidth of a filter is usually measured with the beam normal to the filter (i.e., $\theta = 0°$). Therefore, the half bandwidth $(\Delta\lambda)_\varepsilon$, in angstroms, is given by

$$(\Delta\lambda)_\varepsilon = \frac{\varepsilon}{\beta + \zeta}(\Delta\lambda) = \frac{\lambda\theta_{\max}{}^2}{2t}. \qquad (19)$$

IV. NUMBER OF CHANNELS AVAILABLE FOR COLOR CODING

Three values T, T', and $\theta_{\max}$ must be specified before the number of channels available for color coding can be determined. T and T' are determined by the required signal-to-noise ratio of the over-all system which has not yet been specified. However, typical requirements of the color coding system may be that the intensity of each channel may vary by no more than a factor of two in any beam position and that the crosstalk between each channel be no more than 10 db. This requires that $T = 0.5$ and $T' = 0.05$.

The angle of incidence θ of the beam at the filter is related to the bit capacity N^2 of the digital light deflector. The number of resolvable bits in one direction is given by[2]

$$N = K\theta D/\lambda \qquad (20)$$

where θ is the maximum deflection angle of the beam in one plane, D is the effective aperture of the digital light deflector, λ is the wavelength of the light beam, and K is a constant. The value of K is determined by the transverse mode of propagation of the light beam and the ratio of the linear bit separation in the memory plane compared to the theoretical limit; its value is about 0.2. Practical values are $N = 512$ for $\theta = 4°$ with an aperture diameter of about 2 cms. Present considerations of the digital light deflector have the arbitrary requirement that the center ray of the reflected beam from the memory plane can just be accepted by the aperture of the optical system when the beam deflection is a maximum along the x or y axis. This is shown schematically in Fig. 6. Therefore, the maximum angle of incidence of the light at the filter is the angle Ψ as determined by the extreme rays of the light cone when the beam is focused in the center of the memory plane. To a close approximation, Ψ equals 2θ. The angle of the beam inside the filter is reduced due to refraction at the surface of the dielectric layer in the filter and for small angles $\theta_{\max} = 2\theta/\mu$. The total number of channels M available for color coding is, therefore,

$$M = \frac{B}{(\Delta\lambda)} = \frac{B\mu^2 t}{\lambda\theta^2(t' + t)} \qquad (21)$$

where B is the bandwidth of the modulator.

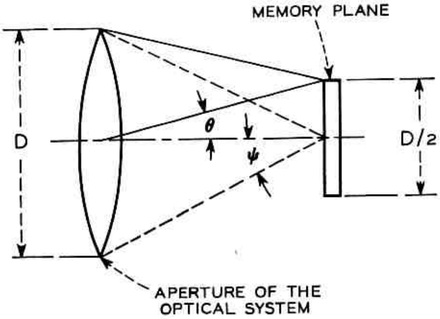


Fig. 6 — Schematic showing the maximal angle of incidence of the beam at the memory plane.

The total capacity C of the digital light deflector and the color coding system is

$$C = N^2M = (K^2B)\frac{D^2\mu^2}{\lambda^3}\frac{t}{(t' + t)}. \tag{22}$$

This relation shows that the total capacity of the system is independent of the maximum beam deflection angle but strongly dependent upon the wavelength of the beam.

The maximum capacity of the color coding system for a beam deflection angle of $\theta = 4°$, also $T = 0.5$, $T' = 0.05$, $D = 2$ cm, $\lambda = 5000$ Å, $\mu = 2.45$ (which is the refractive index for TiO_2 that is frequently used in Fabry-Perot filters), and an optical switch bandwidth of $B = 436$ Å (which is the maximum value for a zero biased KTN switch and a light leakage of $\delta = 0.01$) is 27 channels. This means that the capacity of the digital light deflector that has $N = 512$ resolvable spots in each direction could be increased from 2.6×10^5 to 7.1×10^6 bits. The half bandwidth of the filters for the above values is given by (19) as 12.6 Å, which is well within the capabilities of present day technology.

Unfortunately, when a KTN optical switch is operated with zero electrical bias, the required voltage and power dissipation are greater than can be accepted. The problem is the increase in the light leakage through the switch caused by the change in the operating temperature due to this increase in the dissipated power. A compromise may be possible, however, by using less than 27 color coding channels, hence reducing the required bandwidth, and increasing the allowed light leakage due to color coding. As an example, if the system has 10 color coding channels and the light leakage δ is increased to 0.035, then the optical switches can be biased to $N = 4$; this reduces the dissipated power in the KTN crystal by about a factor of four compared with the unbiased case.

V. CONCLUSIONS

The calculations show that the capacity of the digital light deflector can be increased by a factor of twenty or more by the addition of a color coding scheme. The device could be constructed with present technology by making the optical switches of the digital light deflector from KTN crystals and by using narrow-band transmission filters for the channel selection, but it would have several practical problems. One problem is the necessity to operate the KTN optical switches at zero, or a very low, electrical bias. This would require higher operating voltages and greater power dissipation in the KTN crystals than is anticipated in the present design of the digital light deflector. Other electro-optic materials have been considered for use as the optical switch, but either the angular aperture of the material is too small, the power requirement is too large, or the material is not presently available in large enough pieces. An alternative is to use a stressed optical plate in which the birefringence necessary to produce the 90° rotation of the plane of polarization is induced by applying a mechanical stress. The wavelength bandwidth of a fused quartz stressed plate is approximately 600 Å when operated with a light beam with a wavelength at 5000 Å. This bandwidth can accommodate 37 color coding channels. The disadvantage of the stressed plate optical switch is that it is limited to switching times of the order of milliseconds compared to microseconds for a KTN optical switch.

A disadvantage of using narrow-band transmission filters for channel selection is the large attenuation in the reflected beam due to the achromatic beam divider. This can be overcome by the use of narrow-band reflection filters.

One other item that the color coding system requires is a multiple frequency light source. The light beam should have high intensity, be highly directional, and have a frequency spread between each line equal to the channel separation. One possibility is to use a laser that oscillates at many frequencies simultaneously, such as the argon laser. Another possibility is to use a stimulated Raman source in which several frequencies are produced by down shifting the incident laser beam by Raman scattering from a suitable medium.

VI. ACKNOWLEDGMENTS

It is a pleasure to acknowledge the discussions with J. E. Geusic, who was jointly responsible with the author for initiating this device consideration.

REFERENCES

1. Schmidt, U. J., *Optical Processing of Information*, Spartan Books, Inc., Baltimore, 1963, p. 98.
Nelson, T. J., B.S.T.J. *43*, May 1964, p. 821–845.
Tabor, W. J., B.S.T.J. *43*, May 1964, p. 1153–1154.
Kulcke, W., Harris, T. J., Kosanke, K., and Max, E., IBM J. Res. and Dev. *8* 1964, p. 64–67.
Soref, R. A. and McMahon, D. H., Appl. Opt. *5*, March 1966, p. 425–434.
2. Tabor, W. J., A High-Capacity Digital Light Deflector Using Wollaston Prisms, To be published in B.S.T.J.
3. Geusic, J. E., Kurtz, S. K., Nelson, T. J., and Wemple, S. H., Appl. Phys. Letters, *2*, 1963, p. 185.
4. Chen, F. S., Geusic, J. E., Kurtz, S. K., Skinner, J. G., and Wemple, S. H., to be published in Appl. Phys., January, 1966.
5. Born, M. and Wolf, W., *Principles of Optics*, Pergamon Press, 1959.

A Model for the Random Video Process

By L. E. FRANKS

(Manuscript received December 27, 1965)

For problems concerning the transmission of video signals, it is often desirable to know the statistical distribution of power in the frequency domain for the signal process. It is convenient to have a model, involving only a few essential parameters, which will satisfactorily characterize the power spectral density of the random video signal. This paper proposes a model for the random picture and derives expressions for second-order statistical properties of the video signal obtained from a conventional scanning operation on the picture. The properties of typical picture material make valid certain approximations which lead to especially simple, closed-form expressions for power spectral density. The continuous part of the power spectral density is expressed as a product of three factors, characterizing separately the influence of point-to-point, line-to-line, and frame-to-frame correlation. For parameters representative of typical picture material there is observed an extreme concentration of power near multiples of the line scan and frame scan rates. An illustrative example of the use of the model in an optimum linear filtering problem is included.

I. INTRODUCTION

This paper provides a detailed development of a simple model for characterizing the statistical properties of a random video signal. The primary concern is the modeling of the power spectral density of the electrical process generated by linear, sequential scanning of a rectangular portion of an infinite, two-dimensional random picture. The spatial and temporal statistical properties of typical picture material allow approximations which lead to a model having an especially simple form, characterized by only a few parameters. The model has a form convenient for the analysis of a variety of signal transmission problems. The validity of the model for these purposes is established by comparing it with results obtained in several independent experimental studies.[1,2,3]

The relationship between the second-order statistics of the random picture and those of the resulting video signal due to line-to-line and frame-to-frame correlation is examined in Section II. Section III considers the composite video signal wherein the picture signal is periodically interrupted and a periodic pattern is inserted for purposes of synchronization and blanking. A model of the random picture process is developed in Section IV. The results are combined in Section V to provide a summary of expressions for the power spectral density of the composite video signal. Section VI is an illustrative example of the use of the model for deriving optimum linear signal processing networks for video signal transmission over a noisy channel. A glossary of symbols is provided in Appendix A.

II. EFFECTS OF THE SCANNING OPERATION

For the first step in the development of the random video signal, we consider a still picture with luminance given by the "stationary" random process, $d(x,y)$; i.e., the two-dimensional autocovariance function for the process can be described by

$$E\,[d(x_1\,,y_1)\,d(x_2\,,y_2)] = \varphi\;(\alpha,\beta), \tag{1}$$

where $\alpha = x_2 - x_1 =$ horizontal displacement

$\beta = y_2 - y_1 =$ vertical displacement.

For convenience in derivation of the equations, we assume that $d(x,y)$ is a zero-mean process. Although physically $d(x,y)$ would be non-negative, it is easier to add in the mean in the final expressions. The video signal, $v(t)$, at the output of an optical scanner moving at constant velocity across the picture is a stationary process with an autocorrelation function simply related to $\varphi(\alpha,\beta)$. In order to avoid the introduction of unnecessary constants, assume that the scanner moves in a horizontal direction at unit velocity and also that output voltage is proportional to luminance with unit conversion gain. Then

$$\varphi_1(\tau) = E[v(t)v(t+\tau)] = \varphi(\tau,0)$$

and

$$E[v(t)] = 0. \tag{2}$$

Actually, the scanner output is normally a nonlinear function of luminance, however, the model developed here has the property that its autocovariance function is changed only by a multiplicative constant when subjected to a zero-memory, nonlinear transformation.* Hence, the model remains valid for any scanner characteristic and any in-

* For the random process described in Section IV, the autocovariance is proportional to the variance of the first-order amplitude probability density function.

stantaneous companding operation to which the video signal might be subjected.

The next step is to account for the effects of line-to-line correlation by considering the sequential scanning of a still picture in the form of an infinite strip with a finite horizontal width, traversed by the scanner in an interval of T seconds. Successive lines are separated in a vertical direction by a distance corresponding to the horizontal travel of the scanner in an interval of T_e seconds. The abrupt change in scanner position when it reaches the edge of the strip causes the video signal to be a nonstationary process. The autocorrelation for the process, $\psi(t,\tau) = E[v(t)v(t+\tau)]$ is periodic T in t. This is related to a stationary process in the usual manner by considering t a random variable uniformly distributed over the interval $(0,T]$. Then,

$$\begin{aligned} \varphi_2(\tau) &= E[\psi(t,\tau)] \\ &= \sum_{k=-\infty}^{\infty} \varphi(\tau - kT, kT_e)P(k,\tau). \end{aligned} \tag{3}$$

The probability, $P(k,\tau)$, that the points t and $t+\tau$ fall in lines k apart is given by the translates of the triangular function, $q_T(\tau)$.

$$P(k,\tau) = q_T(\tau - kT),$$

where

$$\begin{aligned} q_T(\tau) &= 1 - \frac{|\tau|}{T} \quad \text{for} \quad |\tau| \leqq T \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{4}$$

Combining (3) and (4), the autocorrelation function for the video signal with line-to-line correlation taken into account becomes

$$\varphi_2(\tau) = \sum_{k=-\infty}^{\infty} \varphi(\tau - kT, kT_e)q_T(\tau - kT), \tag{5}$$

For typical picture material, $\varphi(T/2, \beta) \cong 0$. In this case $\varphi_2(\tau)$ is a sequence of essentially isolated pulse shapes, $q_T(\tau)\varphi(\tau,0)$, centered on integral multiples of T, the kth pulse from the origin having a magnitude proportional to $\varphi(0,kT_e)$ as indicated in Fig. 1.

The power spectral density of this process is

$$\begin{aligned} \Phi_2(f) &= \int_{-\infty}^{\infty} \varphi_2(\tau)e^{-j2\pi f\tau}d\tau \\ &= \sum_{k=-\infty}^{\infty} G(f,kT_e)e^{-j2\pi kTf} \end{aligned} \tag{6}$$

where

$$G(f,kT_e) = \int_{-\infty}^{\infty} q_T(\tau)\varphi(\tau,kT_e)e^{-j2\pi f\tau}d\tau. \tag{7}$$

If we let

$$H(f,\nu) = \int_{-\infty}^{\infty} G(f,\sigma)e^{-j2\pi\nu\sigma}d\sigma \tag{8}$$

then

$$G(f,kT_e) = \int_{-\infty}^{\infty} H(f,\nu)e^{j2\pi kT_e\nu}d\nu. \tag{9}$$

Now substituting (9) into (6) and using the identity

$$\sum_{k=-\infty}^{\infty} e^{-j2\pi kx} = \sum_{m=-\infty}^{\infty} \delta(x - m) \tag{10}$$

we get

$$\Phi_2(f) = \frac{1}{T_e}\sum_{m=-\infty}^{\infty} H\left(f, \frac{T}{T_e}\left(f - \frac{m}{T}\right)\right). \tag{11}$$

Again considering typical picture material, $\varphi(\tau,0)$ is narrow compared to $q_T(\tau)$ so a good approximation is $q_T(\tau)\varphi(\tau,0) \cong \varphi(\tau,0)$. In this case, $H(f,\nu)$ is essentially just the double Fourier transform of the picture autocovariance function

$$H(f,\nu) \cong \iint_{-\infty}^{\infty} \varphi(\tau,\sigma)e^{-j2\pi(f\tau+\nu\sigma)}d\tau d\sigma. \tag{12}$$

A more significant consequence of the correlation in typical picture material is that $\Phi_2(f)$ can be closely approximated by the product of

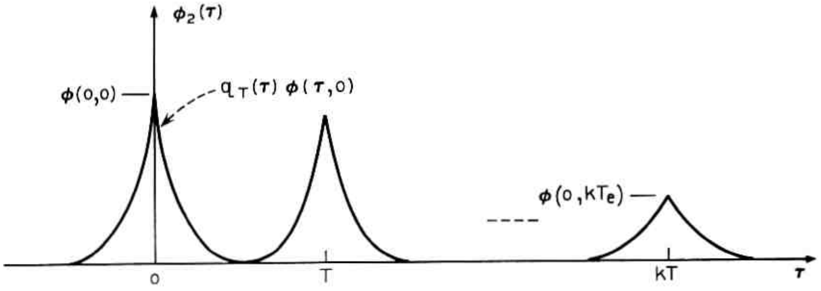


Fig. 1 — Autocorrelation of video signal with line-to-line correlation.

two functions, one periodic $1/T$ and the other an envelope with relatively small variation over $1/T$ intervals. Since, for typical picture material, $H(f,0)$ and $H(0,f)$ have roughly the same width and since $T/T_e \gg 1$, it follows that $H(0,(Tf/T_e))$ is very narrow compared to $H(f,0)$. Furthermore, since $H(f,0)$ has relatively small variation over an interval of width $1/T$, then (11) can be approximated by

$$\begin{aligned}\Phi_2(f) &\cong \frac{1}{T_e}\sum_{m=-\infty}^{\infty} H\left(\frac{m}{T},\frac{T}{T_e}\left(f-\frac{m}{T}\right)\right)\\ &\cong \frac{1}{T_e}H(f,0)\sum_{m=-\infty}^{\infty} H\left(0,\frac{T}{T_e}\left(f-\frac{m}{T}\right)\right).\end{aligned} \tag{13}$$

A power spectral density of the form indicated in (13) corresponds to the property of separability in $\varphi(\tau,\sigma)$. Let

$$\varphi(\tau,\sigma) = \overline{d^2}\,\varphi_h(\tau)\,\varphi_v(\sigma) \tag{14}$$

where $\varphi_h(\tau)$ and $\varphi_v(\sigma)$ are normalized autocorrelation functions for scanning along horizontal and vertical lines, respectively. $\varphi_h(0) = \varphi_v(0) = 1$; $\overline{d^2} = \varphi(0,0)$. Neglecting edge effects, from (12) we have

$$H(f,\nu) = \overline{d^2}\,\Phi_h(f)\,\Phi_v(\nu),$$

and the power spectral density (11) becomes

$$\Phi_2(f) = \frac{\overline{d^2}}{T_e}\Phi_h(f)\sum_{m=-\infty}^{\infty}\Phi_v\left(\frac{T}{T_e}\left(f-\frac{m}{T}\right)\right). \tag{15}$$

A sketch of this function is shown in Fig. 2.

The final step in characterizing the effects of the scanning operation

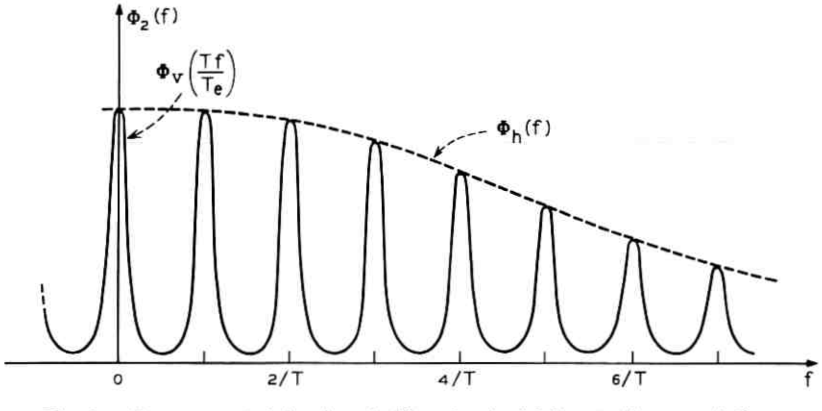


Fig. 2 — Power spectral density of video signal with line-to-line correlation.

is to account for frame-to-frame correlation in repeated scanning of a finite portion of the infinite strip. Of course, if the picture is still, the resulting process is periodic. We consider a randomly moving picture with slow variation compared to the frame repetition rate, $1/NT$, where vertical scanning is accomplished by N uniformly spaced lines. The nonstationarity arising from the abrupt change of scanner position when it reaches the bottom edge of the picture is handled in the same manner as before in terms of a shaping function, $q_{NT}(\tau)$, whose effect can be neglected for typical picture material. Expressing the normalized correlation of the luminance of a picture point at times separated by k frame intervals by $\varphi_t(kNT)$; $\varphi_t(0) = 1$, then

$$\varphi_3(\tau) = \sum_{k=-\infty}^{\infty} \varphi_t(kNT)\varphi_2(\tau - kNT). \tag{16}$$

Because of the slow variation due to motion, frame-to-frame correlation is high and $\varphi_3(\tau)$ is essentially the product of $\varphi_t(\tau)$ and a periodic repetition of $\varphi_2(\tau)$ as indicated in Fig. 3.

The power spectral density, obtained from (16) and the use of relation (10) is

$$\begin{aligned}\Phi_3(f) &= \Phi_2(f) \sum_{k=-\infty}^{\infty} \varphi_t(kNT)e^{-j2\pi kNTf} \\ &= \frac{1}{NT}\,\Phi_2(f) \sum_{m=-\infty}^{\infty} \Phi_t\left(f - \frac{m}{NT}\right).\end{aligned} \tag{17}$$

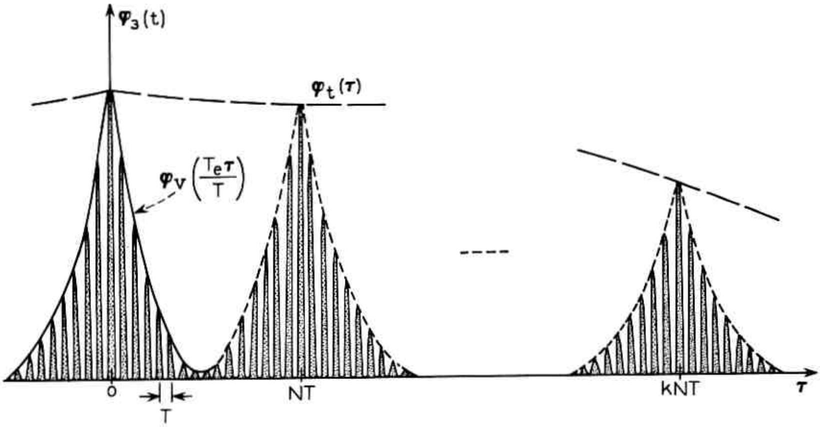


Fig. 3 — Autocorrelation of video signal with frame-to-frame correlation.

Combining (17) and (15), the final expression for $\Phi_3(f)$ is given as the product of three functions; an envelope $G_h(f)$ representing horizontal picture correlation, a function $G_v(f)$, periodic $1/T$, representing vertical picture correlation, and a function $G_t(f)$, periodic $1/NT$, representing frame-to-frame correlation.

$$\begin{aligned}\Phi_3(f) &= [\overline{d^2}\Phi_h(f)]\left[\frac{1}{T_e}\sum_{m=-\infty}^{\infty}\Phi_v\left(\frac{T}{T_e}\left(f-\frac{m}{T}\right)\right)\right] \\ &\qquad\cdot\left[\frac{1}{NT}\sum_{l=-\infty}^{\infty}\Phi_t\left(f-\frac{l}{NT}\right)\right] \qquad (18) \\ &= G_h(f)G_v(f)G_t(f).\end{aligned}$$

As indicated in Fig. 4, the factors $G_v(f)$ and $G_t(f)$ impose a "fine structure" on $\Phi_3(f)$. In considering various smoothed versions of power spectral density, it is helpful to note that the average values of $G_v(f)$ and $G_t(f)$ are both unity.

Some practical scanning operations involve line interlacing. For the conventional 2:1 interlace scan, the resulting modification of (18) is simple. Since consecutive lines are now twice as far apart, the factor $G_v(f)$ is modified by replacing T_e with $2T_e$. This modification results in the individual peaks in $G_v(f)$ centered at multiples of $1/T$ being broadened to twice their original width. Since the picture is scanned vertically every $NT/2$ seconds, the $G_t(f)$ factor is modified by replacing N by $N/2$. This effects a suppression of the terms centered at odd

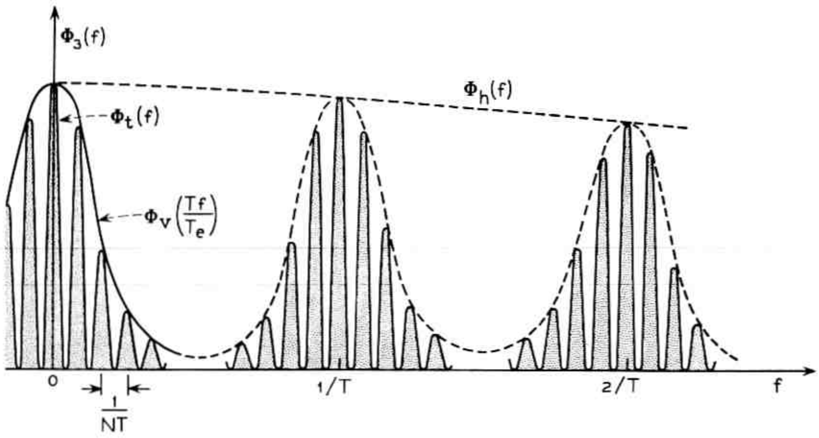


Fig. 4 — Power spectral density of video signal with frame-to-frame correlation.

multiples of $1/NT$. This latter modification is approximate since consecutive vertical scans are not exactly in register, however the ratio of power at odd multiples of $1/NT$ to power at even multiples of $1/NT$ is given by the ratio of $1 - \varphi_v(T_e)$ to $1 + \varphi_v(T_e)$ which for typical picture material may be less than 0.01. Derivation of the modifications for more complicated interlacing arrangements is straightforward.

III. COMPOSITE VIDEO SIGNAL

It is common practice to interrupt the video signal after each line scan and frame scan for the purpose of inserting control signals such as synchronizing and blanking pulses. The resulting signal is referred to as the composite video signal and the following development shows the form of the power spectral density. In order to simplify the argument, consider the following composite signal, $z(t)$, which is a random process interrupted every T seconds for a duration of αT seconds with an arbitrary periodic pattern, $w(t)$, inserted in the blank interval:

$$z(t) = p(t)\, v(t) + w(t) \tag{19}$$

where $p(t)$ is periodic T, equal to zero in the blank interval and equal to one in the video interval; $w(t)$ is periodic T and equal to zero in the video interval; and $v(t)$ is a zero-mean random process with autocorrelation, $\varphi(\tau)$. The process $p(t)\ v(t)$ is, of course, nonstationary but after averaging its autocorrelation function, $\psi(t,\tau)$, over the period T we have

$$\begin{aligned} \hat{\varphi}(\tau) &= E\,[\psi(t,\tau)] \\ &= \varphi(\tau) \text{ Pr } [t \text{ and } t + \tau \text{ in video region}] \\ &= (1 - \alpha)\varphi(\tau) \sum_{k=-\infty}^{\infty} q_{(1-\alpha)T}(\tau - kT) \end{aligned} \tag{20}$$

where

$$\begin{aligned} q_{(1-\alpha)T}(\tau) &= 1 - \frac{|\tau|}{(1 - \alpha)T} \quad \text{in} \quad |\tau| \leqq (1 - \alpha)T \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

From (20), the autocorrelation function of the periodically blanked process is obtained by multiplying the autocorrelation of the original process by a periodic function having a shape as indicated in Fig. 5. Note that no periodic components are generated by blanking. For moderately small α and typical video signal autocorrelation functions,

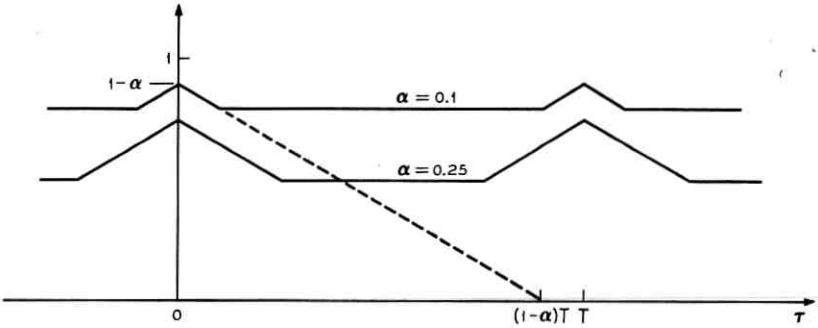


Fig. 5 — Periodic shaping for autocorrelation function of periodically blanked video process.

the effect of multiplying by the periodic shaping function is essentially the same as multiplying by the constant, $1 - \alpha$. The power in the composite signal is the sum of the power in the blanked process and the power in the added periodic signal since $p(t)v(t)w(t) \equiv 0$. Hence, the power spectral density for the composite signal is

$$\Phi_z(f) = (1 - \alpha)\Phi(f) + \sum_l |w_l|^2 \delta\left(f - \frac{l}{T}\right) \tag{21}$$

where the w_l are the Fourier coefficients for $w(t)$. Modification of (21) for the actual control signal which consists of both horizontal and vertical synchronizing and blanking pulses is straightforward. In this case, $w(t)$ is replaced by a signal periodic NT (or $NT/2$ for 2:1 interlace). The constant $1 - \alpha$ is still just the relative amount of time devoted to the video signal.

IV. MODEL FOR RANDOM PICTURE

To complete the model for the random video signal it remains to characterize the luminance process, $d(x,y)$, in such a manner that a useful expression for its autocovariance, $\varphi(\alpha,\beta)$, can be derived. The discussion in Section II indicated the validity of the separable form (14) for $\varphi(\alpha,\beta)$. Assuming separability, we need only model the one-dimensional process resulting from a unit velocity, linear scanning of the picture. Assume that a realization of this process is a piecewise-constant function, $v(t)$, which takes on the value v_n over the interval, $t_n \leqq t < t_{n+1}$ as shown in Fig. 6. The occurrence of the sequence $\{t_n\}$ of points along the t-axis is a stationary random process and the sequence

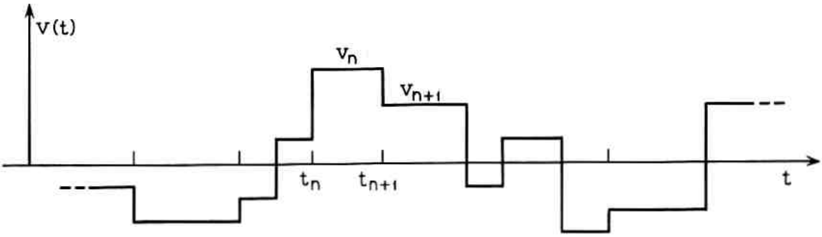


Fig. 6 — Random video signal.

$\{v_n\}$ of random variables is also stationary. Then the autocorrelation function for $v(t)$ is

$$\varphi(\tau) = \overline{d^2} \sum_{m=0}^{\infty} r_m P(m,\tau) \tag{22}$$

where

$$\overline{d^2}\, r_m = E\,[v_n v_{n+m}]$$

$$\overline{d^2} = E\,[v_n^{\,2}]; \qquad E\,[v_n] = 0$$

and $P(m,\tau)$ is the probability that the points t and $t + \tau$ are in intervals m apart, i.e., that m points of the $\{t_n\}$ sequence lie between them.

The simplest model is constructed by assuming that the $\{v_n\}$ are statistically independent and that the $\{t_n\}$ are generated by a Poisson process with rate parameter, λ. This model is the random step function discussed by Laning and Battin.[5] For this case, (22) reduces to

$$\varphi(\tau) = \overline{d^2}\, P(0,\tau) \tag{23}$$

with

$$P(m,\tau) = \frac{(\lambda\,|\,\tau\,|)^m}{m!}\, e^{-\lambda|\tau|}$$

hence,

$$\varphi(\tau) = \overline{d^2}\, e^{-\lambda|\tau|}\,. \tag{24}$$

An obvious step in the generalization of this model is to consider correlation in the $\{v_n\}$ sequence. Suppose $\{v_n\}$ is a stationary, wide-sense Markoff sequence.[6] Then it has the property that

$$r_m = r_1^{\,m} \tag{25}$$

where r_1 is the correlation between adjacent elements of the sequence. In this case,

$$\begin{aligned}\varphi(\tau) &= \overline{d^2} \sum_{m=0}^{\infty} \frac{(\lambda \mid \tau \mid)^m}{m!} r_1^m e^{-\lambda|\tau|} \\ &= \overline{d^2} \exp\left[-(1 - r_1)\lambda \mid \tau \mid\right],\end{aligned} \tag{26}$$

hence, this model is equivalent to the previous one (24) with λ replaced by $(1 - r_1)\lambda$. This result suggests an interesting alternate formulation of the model. Suppose that the $\{t_n\}$ sequence is uniformly spaced with separation T_e (stationarity is accomplished by randomizing the phase of the sequence). For this case,

$$P(m,\tau) = q_{T_e}(\tau - mT_e) + q_{T_e}(\tau + mT_e) \tag{27}$$

where

$$\begin{aligned} q_{T_e}(\tau) &= 1 - \frac{\mid \tau \mid}{T_e} \qquad \text{in} \quad \mid \tau \mid \leqq T_e \\ &= 0 \qquad \text{otherwise.}\end{aligned}$$

Let $\{v_n\}$ be a stationary, wide-sense Markoff sequence with correlation ρ between adjacent elements, then

$$\begin{aligned}\varphi(\tau) &= \overline{d^2} \sum_{l=-\infty}^{\infty} \rho^{|l|} q_{T_e}(\tau - lT_e) \\ &\cong \overline{d^2} \exp\left(-\hat{\lambda} \mid \tau \mid\right) \qquad \text{where } \hat{\lambda} = -\frac{\ln \rho}{T_e}.\end{aligned} \tag{28}$$

This is a polygonal approximation to the exponential function which, since correlation between points T_e apart is typically very large, is a very close approximation.

Using the preceding models, which are all equivalent, the random picture is characterized by an autocovariance

$$\varphi(\tau,\sigma) = \overline{d^2} \exp\left(-\lambda_h \mid \tau \mid - \lambda_v \mid \sigma \mid\right) \tag{29}$$

which depends only on the variance, $\overline{d^2}$, of luminance and two parameters λ_h and λ_v which specify the average number of statistically independent luminance levels in a unit distance along the horizontal and vertical, respectively. Alternatively, the correlation is characterized by the parameters $\rho_h = \exp[-\lambda_h T_e]$ and $\rho_v = \exp[-\lambda_v T_e]$ which are the correlation coefficients of luminance in adjacent picture elements when the picture area is quantized into small squares of dimension T_e.

The suitability of the exponential correlation function for modeling the random picture can be established by examining the results of correlation measurements reported by Kretzmer[1] and O'Neal[2] and power spectral density measurements by Deriugin.[3]

V. SUMMARY

Combining the results of the preceding sections, simple closed-form expressions can be written for the power spectral density of the composite video signal. At this point the mean value $\bar{d}$ of the luminance is included so the variance of the luminance becomes $\overline{d^2} - \bar{d}^2$. The synchronizing and blanking signals are assumed to occupy a fraction α of the total time and to form a pattern periodic NT. Let w_l be the Fourier coefficients of the periodic signal added to $\bar{d}$ in the blank intervals. Then the power spectral density for the composite signal is

$$\begin{aligned} S(f) &= (1 - \alpha) G_h(f) G_v(f) G_t(f) \\ &\quad + \sum_{l=-\infty}^{\infty} | w_l |^2 \, \delta\left(f - \frac{l}{NT}\right) + \bar{d}^2 \, \delta(f). \end{aligned} \tag{30}$$

Using the exponential correlation functions of Section IV and (18), we have

$$G_h(f) = (\overline{d^2} - \bar{d}^2) \frac{2\lambda_h}{(2\pi f)^2 + \lambda_h^2}. \tag{31}$$

The factor indicating shaping due to line-to-line correlation becomes

$$G_v(f) = \frac{1}{T_e} \sum_{m=-\infty}^{\infty} \frac{2\lambda_v}{\left[2\pi T/T_e\left(f - \frac{m}{T}\right)\right]^2 + \lambda_v^2}. \tag{32}$$

The closed form for this expression is obtained by noting that $G_v(f)$ is the convolution product,

$$G_v(f) = \frac{T}{T_e} \frac{2\lambda_v}{(2\pi Tf/T_e)^2 + \lambda_v^2} * \frac{1}{T} \sum_m \delta\left(f - \frac{m}{T}\right). \tag{33}$$

Using the identity (10) and performing the indicated convolution, $G_v(f)$ is expressed as a geometric series which can be summed to give

$$G_v(f) = \frac{\sinh T_e\lambda_v}{\cosh T_e\lambda_v - \cos 2\pi Tf}. \tag{34}$$

A similar expression for the $G_t(f)$ factor can be obtained by assuming that luminance of a point at successive frames forms a wide-sense Markoff sequence with frame-to-frame correlation $\rho_t = \exp\,[-\lambda_t NT]$.

Then

$$G_t(f) = \frac{\sinh NT\lambda_t}{\cosh NT\lambda_t - \cos 2\pi NTf}. \tag{35}$$

Frame-to-frame correlation, ρ_t, has been measured experimentally[1] and found to lie between 0.86 and 0.80 for typical material. Using these values in (35) the width (between -3-db points) of the peaks in $G_t(f)$ is only about 0.048 to 0.071 of the separation, $1/NT$, between the peaks.

In applications where the structure of $G_t(f)$ is too fine to resolve, the smoothed version, $(1 - \alpha)G_h(f)G_v(f)$, of the continuous part of the power spectral density is of interest. This quantity is shown in Fig. 7 for two different types of video signal; one the standard 525-line, 30-frame per second broadcast television signal (BCTV) and the other the 275-line, 30-frame per second, initial design of the Bell System station-to-station *Picturephone*® service (PP). For comparison, it is assumed that both pictures have the same correlation, $\rho = 0.9$, between picture elements of dimension T_e in both horizontal and vertical directions. Also both signals are obtained from 2:1 interlaced scanning. The difference in the two curves in Fig. 7 is due to different values of the quantity, T_e/T. T_e/T is a fundamental parameter of the raster design depending

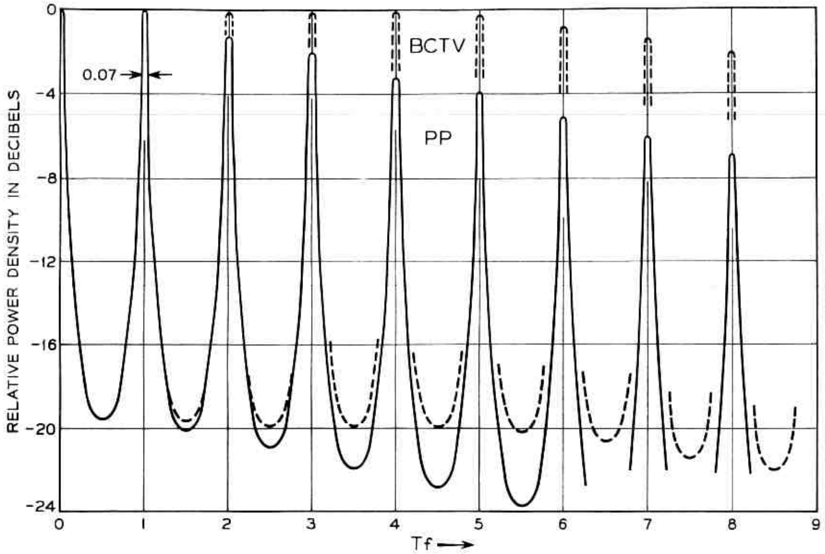


Fig. 7 — Continuous part of power spectral density for two typical video signals with frame rate structure smoothed out. ($\rho = \rho_h = \rho_v = 0.9$.)

on the number of lines per frame, the aspect ratio of the visible portion of the picture, and the relative size of the blank portions of the horizontal and vertical scans. For BCTV, $T_e/T = 0.00128$; and for PP, $T_e/T = 0.0041$.

The value of the parameter ρ used in Fig. 7 represents a highly detailed random picture. A typical head-and-shoulders view of a person may have $\rho = 0.99$ and $\rho = 0.98$ represents a moderately detailed picture.[1] Even for $\rho = 0.9$ the power is extremely concentrated around multiples of the line scan rate, $1/T$. The width of the peaks between -3-db points and also the ratio between successive maxima and minima in the power density are shown in Table I for various values of ρ and λT_e.

VI. APPLICATION

One obvious application of the model is in problems concerning minimum mean-squared error filtering of the video signal in noise. Solution of these problems invariably requires a knowledge of power spectral densities of the signal and noise. The concept of utilizing the inherent redundancy in the video signal to ease transmission requirements is familiar. Alternative to seeking coding arrangements which reduce bandwidth requirements, we can consider linear processing operations which utilize the highly nonuniform nature of the power spectral density to effect a reduction of signal power needed for a given performance.

As an example of this approach, consider the design of optimum pre-emphasis and de-emphasis filters for transmission over a noisy channel as shown in Fig. 8. Assume that the channel has a constraint on maximum signal power and that it is desired to minimize mean-squared error in the received signal, $y(t)$.

Details of the derivation of the optimum filtering characteristics

TABLE I — WIDTH AND HEIGHT OF POWER DENSITY CONCENTRATIONS ABOUT MULTIPLES OF THE LINE SCANNING RATE FOR 2:1 INTERLACED SCANNING

ρ	$\lambda T_e = -\ln \rho$	Width (between -3 db Points) Relative to $1/T$	$10 \log_{10} \frac{S(n/T)}{S(n + \frac{1}{2}/T)}$
0.99	0.010	0.00636	40.0 db.
0.98	0.020	0.0127	34.0
0.97	0.030	0.0191	30.4
0.95	0.051	0.0326	25.8
0.90	0.105	0.0668	19.6
0.85	0.163	0.1035	15.8
0.80	0.223	0.1420	13.2

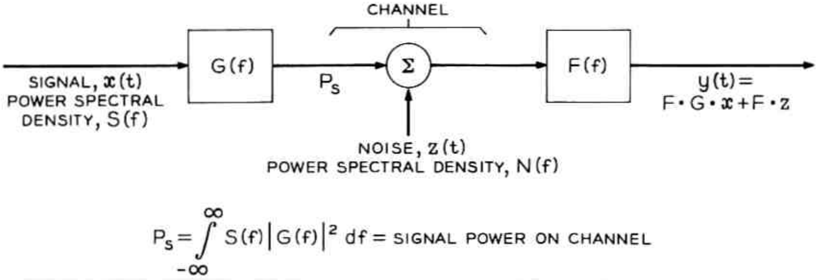

CHANNEL
G(f)
Σ
F(f)
SIGNAL, x(t)
POWER SPECTRAL
DENSITY, S(f)
P_S
y(t)=
F•G•x+F•z
NOISE, z(t)
POWER SPECTRAL DENSITY, N(f)
P_S = ∫_{-∞}^{∞} S(f)|G(f)|² df = SIGNAL POWER ON CHANNEL

Fig. 8 — Spectrum shaping for transmission over noisy channel.

are presented in Appendix B. For high signal-to-noise ratio on the channel, the two filters are essentially inverse and we get*

$$F(f) = G^{-1}(f) = \left[\frac{\mu S(f)}{N(f)}\right]^{\frac{1}{4}} \tag{36}$$

where the constant μ is adjusted to meet the signal power constraint.

$$\mu^{\frac{1}{2}} = \frac{1}{P_s}\int (SN)^{\frac{1}{2}}\, df. \tag{37}$$

The advantage gained by using the filter networks can be expressed as a signal-to-noise ratio improvement factor γ which is simply the ratio of signal-to-noise ratios at the output of the receiver with and without filtering.

$$\gamma = \frac{\int S df \int N df}{\int SF^{-2} df \int NF^2 df}\cdot \tag{38}$$

When the optimum filter pair (36) is used, (38) becomes

$$\gamma_{\text{opt}} = \frac{\int S df \int N df}{[\int (SN)^{\frac{1}{2}} df]^2}\cdot \tag{39}$$

The expression for γ_{opt} provides another interesting measure of the nonuniformity of $S(f)$. In a function space representation, the quantity, $\cos^{-1}\ (1/\gamma_{\text{opt}})^{\frac{1}{2}}$, is conventionally interpreted as the angle between the functions $S^{\frac{1}{2}}(f)$ and $N^{\frac{1}{2}}(f)$. If we consider flat noise over a band W and zero outside, then γ_{opt} becomes a comparison of $S^{\frac{1}{2}}(f)$ with a con-

* The expressions for $S(f)$ use only the continuous part of the expression. The discrete componenets should not be transmitted in this problem.

stant over the band of interest. In this case,

$$\gamma_{\text{opt}} = \frac{2W \int S df}{\left[\int_{-W}^{W} S^{\frac{1}{2}} \, df \right]^2}. \tag{40}$$

Values of γ_{opt} in (40) are plotted in Fig. 9 for the case, $W = 60/T$ and $T_e/T = 0.0041$; these parameters corresponding to the *Picturephone* video signal.

The form of the optimum filter, in this case $F(f) = S^{\frac{1}{4}}(f)$, suggests a rather difficult realization problem. Accordingly, it is interesting to evaluate the performance of a suboptimum filter pair having the simple form shown in Fig. 10. The parameter, a, in the network realizing the periodic part of the transfer function is adjusted to match the maxima and minima in $S^{\frac{1}{4}}(f)$. The values of γ using this filter pair are also shown on Fig. 9 and are seen to be remarkably close to the optimum values. It is of interest to note that part of the pre-emphasis filter effects a smoothing of the transmitted power spectral density by transmitting only the partial difference, $x(t) - ax(t - T)$, between successive lines. This technique has been discussed by Harrison[7] and O'Neal[2] with regard

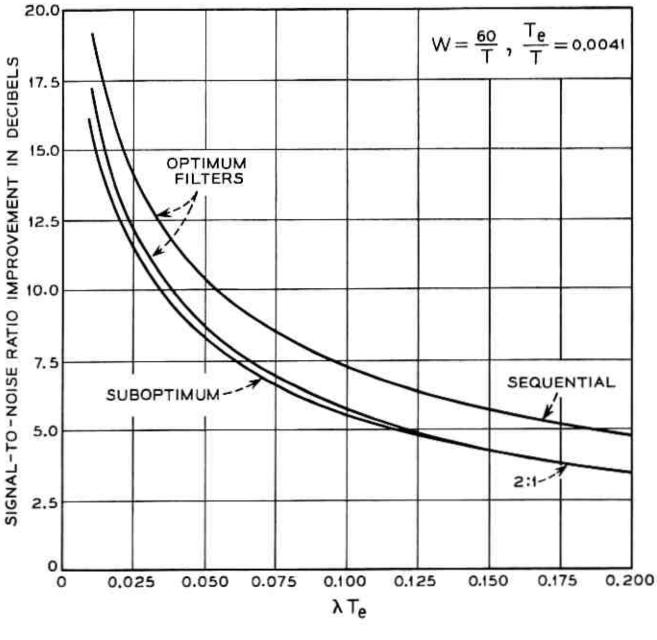
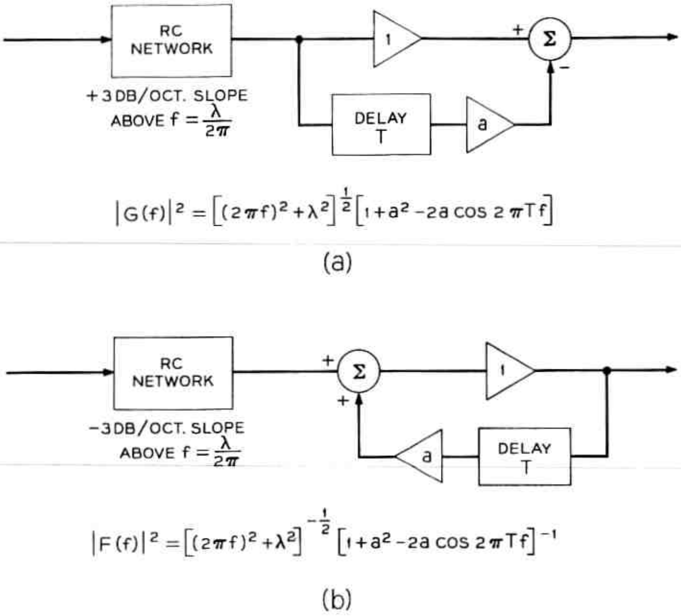


Fig. 9 — Performance of spectrum shaping filters. ($W = 60/T$, $T_e/T = 0.0041$.)


RC NETWORK
+3 DB/OCT. SLOPE ABOVE f = λ/2π
1
+
Σ
-
DELAY T
a
(a)
RC NETWORK
-3 DB/OCT. SLOPE ABOVE f = λ/2π
+
Σ
+
1
a
DELAY T
(b)

$$|G(f)|^2 = \left[(2\pi f)^2 + \lambda^2\right]^{\frac{1}{2}}\left[1 + a^2 - 2a\cos 2\pi Tf\right]$$

$$|F(f)|^2 = \left[(2\pi f)^2 + \lambda^2\right]^{-\frac{1}{2}}\left[1 + a^2 - 2a\cos 2\pi Tf\right]^{-1}$$

Fig. 10 — Suboptimum filter pair; (a) pre-emphasis network, (b) de-emphasis network.

to the use of "previous line" linear prediction to reduce the correlation present in the transmitted signal.

In actual practice, a substantial portion of the indicated advantage may not be realizable. This is because the received noise, in passing through $F(f)$, is concentrated at multiples of $1/T$ causing a line-to-line correlation which is subjectively more annoying than the same amount of flat noise power. This subjective effect has not yet been fully evaluated. If it can be described by a frequency domain weighting function, then the methods presented in Appendix B can be easily adapted to give a more accurate evaluation of optimum filtering.

APPENDIX A

Glossary of Symbols

$d(x,y)$	luminance at the point (x,y) on the picture.
$\varphi(\alpha,\beta)$	autocorrelation of picture luminance.
$\varphi_1(\tau)$	autocorrelation of video signal obtained from linear horizontal scanning of infinite picture.

$\varphi_2(\tau)$ autocorrelation of video signal obtained by sequential horizontal scanning of infinite vertical strip.

$\varphi_3(\tau)$ autocorrelation of video signal obtained by repeated scanning of rectangular portion of moving picture.

$\varphi_h(\tau)$, $\varphi_v(\tau)$ normalized versions of $\varphi(\tau,0)$ and $\varphi(0,\tau)$, respectively.

$\varphi_t(\tau)$ normalized autocorrelation of luminance at a point as a function of time.

$G_h(f)$, $G_v(f)$, $G_t(f)$ factors of the power spectral density characterizing point-to-point, line-to-line, and frame-to-frame correlation, respectively.

$S(f)$ power spectral density of composite video signal.

$w(t)$ periodic part of composite video signal less the average luminance, $\bar{d}$.

w_l lth Fourier coefficient of $w(t)$.

α relative amount of time occupied by non-video portion of the signal.

T line scan interval in seconds.

T_e time interval equivalent to distance between adjacent lines at scanner velocity.

N number of lines per frame.

λ_h, λ_v Poisson rate parameter describing luminance process in horizontal and vertical directions, respectively.

ρ_h, ρ_v correlation between luminance values in adjacent square picture elements, of dimension T_e, in horizontal and vertical directions, respectively.

APPENDIX B

Optimum Filtering of Random Video Signals

The criterion for optimum performance is expressed in terms of mean-squared deviation between the received signal, $y(t)$, and the transmitted signal, $x(t)$, shown in Fig. 8. The received signal is decomposed into distorted signal component, $u(t) = FG \cdot x(t)$, and noise component, $v(t) = F \cdot z(t)$.

Let

$$\begin{aligned} I &= E\,[(x - y)^2] \\ &= E\,[(x - u)^2] + E\,[v^2] - 2E\,[(x - u)v]. \end{aligned} \tag{41}$$

The last term in (41) vanishes by assuming statistical independence of signal and noise and zero-mean for the noise. The two remaining functionals are expressed in the frequency domain as

$$I = \int S(f) \mid 1 - F(f)G(f) \mid^2 df + \int N(f) \mid F(f) \mid^2 df. \qquad (42)$$

We want to find the filter transfer functions $F(f)$ and $G(f)$ such that I is minimized subject to the constraint on signal power on the channel.

$$\int S(f) \mid G(f) \mid^2 df = P_s\,. \qquad (43)$$

Problems of this type appear to have been first discussed by Costas.[8] It can be shown that if the signal-to-noise ratio on the channel is moderately high, then $F(f)$ and $G(f)$ are essentially inverse. Accordingly, we add the constraint, $G^{-1}(f) = F(f)$, which makes the first term in (42) vanish, and find the stationary points of $I + \mu P_s$ with respect to F.

$$I + \mu P_s = \int NF^2\, df + \mu \int SF^{-2}\, df \qquad (44)$$

where we assume F to be a real function (since its phase does not affect signal power or noise power) with the understanding that the pre-emphasis filter can have an arbitrary phase shift since the de-emphasis filter has the complementary phase shift. In order that the first variation of the functional in (44) vanish, it is necessary that

$$F^2 = \left(\frac{\mu S}{N}\right)^{\frac{1}{2}} \qquad (45)$$

where

$$\mu^{\frac{1}{2}} = \frac{1}{P_s} \int (SN)^{\frac{1}{2}}\, df$$

in order to meet the constraint on signal power. Substituting (45) into (44) the minimum mean-squared interference becomes

$$I_{\min} = \frac{1}{P_s}\left[\int (SN)^{\frac{1}{2}}\, df\right]^2. \qquad (46)$$

Because of the constraint, $G^{-1} = F$, it makes sense to speak of signal-to-noise ratio at the output of the receiver. The improvement in signal-to-noise ratio by choosing F according to (45) relative to $F \equiv 1$ is expressed as

$$\gamma_{\text{opt}} = \frac{\int S\, df \int N\, df}{\left[\int (SN)^{\frac{1}{2}}\, df\right]^2}. \tag{47}$$

If we assume that $N(f)$ is constant over the band $|f| \leqq W$ and zero elsewhere, then (47) becomes

$$\gamma_{\text{opt}} = \frac{2W \int S\, df}{\left[\int_{-W}^{W} S^{\frac{1}{2}}\, df\right]^2}. \tag{48}$$

Now let $S(f)$, as indicated in Section V with $\lambda_h = \lambda_v = \lambda$, be given by

$$S(f) = K\left[\frac{2\lambda}{(2\pi f)^2 + \lambda^2}\right]\left[\frac{\sinh \lambda T_e}{\cosh \lambda T_e - \cos 2\pi T f}\right]$$

and since

$$\int S(f)\, df = \varphi(0) = K,$$

we have

$$\gamma_{\text{opt}} = 2W\left\{\int_{-W}^{W}\left[\frac{2\lambda}{(2\pi f)^2 + \lambda^2}\right]^{\frac{1}{2}}\left[\frac{\sinh \lambda T_e}{\cosh \lambda T_e - \cos 2\pi T f}\right]^{\frac{1}{2}} df\right\}^{-2}. \tag{49}$$

Integrals of this type, having an integrand $A(f)\, B(f)$ where $B(f)$ is periodic $1/T$ and $A(f)$ is a slowly changing envelope function, can be closely approximated by

$$\int_{-W}^{W} A(f)B(f)\, df \cong \sum_{m=-WT}^{WT} A\left(\frac{m}{T}\right) \int_{1/T} B(f)\, df \tag{50}$$

$$\cong T\int_{-W}^{W} A(f)\, df \int_{1/T} B(f)\, df.$$

Accordingly, we evaluate

$$\int_{-W}^{W}\left[\frac{2\lambda}{(2\pi f)^2 + \lambda^2}\right]^{\frac{1}{2}} df \cong \frac{(2\lambda)^{\frac{1}{2}}}{\pi} \ln \frac{4\pi W}{\lambda} \qquad \text{for} \qquad \frac{W}{\lambda} \geqq 1$$

and using a suitable change of variable on the second integral

$$\int_{1/T}\left[\frac{\sinh \lambda T_e}{\cosh \lambda T_e - \cos 2\pi T f}\right]^{\frac{1}{2}} df = \frac{2}{\pi T}\tanh^{\frac{1}{2}}\frac{\lambda T_e}{2}\int_0^{\pi/2}\frac{d\varphi}{[1 - k^2\sin^2\varphi]^{\frac{1}{2}}}$$

where

$$k^{-1} = \cosh\frac{\lambda T_e}{2}.$$

This last integral is recognized as the complete elliptic integral of the first kind, $\mathcal{K}\ (\sin^{-1} k)$, which can be obtained from tables. Now combining these results,

$$\left[\int_{-W}^{W} S^{\frac{1}{2}}\, df\right]^2 = \frac{8\lambda}{\pi^4}\tanh\frac{\lambda T_e}{2}\left[\ln\frac{4\pi W}{\lambda}\right]^2 \mathcal{K}^2\ (\sin^{-1} k) \qquad (51)$$

and (49) becomes

$$\gamma_{\text{opt}}(\lambda) = \frac{\dfrac{\pi^4 W}{4\lambda}}{\tanh\dfrac{\lambda T_e}{2}\left[\ln\dfrac{4\pi W}{\lambda}\right]\mathcal{K}^2(\sin^{-1} k)}. \qquad (52)$$

This function is plotted in Fig. 9 over the range of typical values of λ.

For the suboptimum filters shown in Fig. 10, we want to evaluate γ in (38) for the flat noise case. The integrals are evaluated using the approximation indicated in (50).

$$\gamma(\lambda) = \frac{\pi^2(W/\lambda)}{\dfrac{1}{1 - a^2}\left[\ln\dfrac{4\pi W}{\lambda}\right]^2[1 + a^2 - 2ae^{-\lambda T_e}]}. \qquad (53)$$

In (53) the parameter a is selected to satisfy

$$(1 - a/1 + a)^2 = \tanh \lambda T_e/2. \qquad (54)$$

This choice of a makes the ratio of maxima to minima in the periodic part of the transfer function equal to that of the optimum filter. Values of $\gamma(\lambda)$ from (53) with T_e replaced by $2T_e$ for 2:1 interlace scanning are also shown on Fig. 9.

REFERENCES

1. Kretzmer, E. R., Statistics of Television Signals, B.S.T.J., *31*, July, 1952, pp. 751–763.
2. O'Neal, J. B., Predictive Quantizing Systems (Differential Pulse Code Modulation) for the Transmission of Television Signals, To be published.

3. Deriugin, N. G., The Power Spectrum and the Correlation Function of the Television Signal, Telecommunications #7, 1952, pp. 1–12.
4. Ignat'yev, N. K., Power Spectrum of a Signal Obtained by Scanning, Radio Eng. Electron. Phys., *6,* No. 1, Jan., 1961, pp. 19–23.
5. Laning, J. H. and Battin, R. H., *Random Processes in Automatic Control,* McGraw-Hill Book Co., New York, 1956, Ch. 3.
6. Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill Book Co., New York, 1965, Ch. 15.
7. Harrison, C. W., Experiments with Linear Prediction in Television, B.S.T.J., *31,* July, 1952, pp. 764–783.
8. Costas, J. P., Coding with Linear Systems, Proc. IRE, *40,* No. 9, Sept., 1952, pp. 1101–1103.

Theory of Cascaded Structures: Lossless Transmission Lines

By B. K. KINARIWALA

(Manuscript received January 12, 1966)

Cascaded structures play a major role in many signal processing and signal propagating systems. The universality of such structures is particularly evident when the signals are of a wave nature, i.e., the components of the structure are representable by transmission lines rather than lumped elements. Transmission media with discontinuities are examples of such structures. Other examples include integrated, microwave, and optical circuits.

Theory of distributed structures has, so far, been successfully developed only for structures whose components are lossless (or RC) transmission lines oj equal electrical lengths. It is the purpose of this paper to present a theory of cascaded structures when the component elements are lossless transmission lines of arbitrary electrical lengths. Extensions of the theory developed here to other structures will be discussed in a subsequent paper.

I. INTRODUCTION

1.1 *Purpose*

A large class of signal processing and signal propagating systems takes the form of a cascade of elementary two-port, linear transducers. For example, in the classical filter theory cascades of constant-k, m-derived sections, etc. and in the modern network synthesis cascades of transmission-zero sections form the conceptual basis. Integrated circuits utilize RC transmission lines in cascade. In microwave filter theory, the structure takes the form of a cascade of quarter-wave transformers. Optical filters incorporate the same idea in multilayer dielectric thin-film structures. In propagation problems, one typically encounters waves (electromagnetic, acoustic, etc.) travelling in cascades of transmission media and discontinuities. These are but a few examples to indicate the importance and universality of such structures.

It is the purpose of this paper to present a theory of such structures when the component two-ports are representable by uniform lossless transmission lines of arbitrary electrical lengths. Extensions of the theory to include other structures as well as lumped network elements will be discussed in a subsequent paper. A secondary aim of this paper is to incorporate into the theory those algorithms for analysis and synthesis that are most appropriate for computing purposes.

The distinguishing feature of this study is the novel formulation for the transmission matrix specifying each transmission line. The total signal quantities at input and output of the line are related to each other in terms of the forward and backward travelling waves in the line. In the past, results have been obtained only for those structures in which the component transmission lines are of equal electrical length. The new formulation presented here leads to a complete theory of lossless lines in cascade. The analysis and synthesis algorithms obtained here are particularly simple and straightforward. They appear to be quite promising for computation.

1.2 *Background*

In the theory of lumped networks, extensive literature exists on cascaded (lumped) structures. The difficulty arises when some or all of the component two-ports consist of distributed elements. In the case of lumped elements, the system functions are defined by rational functions of the complex frequency variable for which there exist many well-known mathematical results. When distributed elements are present, the system functions involve transcedental functions of the complex variable with a consequent increase in complexity. It has been possible in the past to obtain significant results only for certain classes of transmission line structures by applications of Richards' transformation. In particular, Richards[1] showed that distributed structures consisting only of *uniform, lossless* transmission lines of *equal electrical lengths* are equivalent, under a change of variable, to lumped networks. Many techniques and results of the lumped network theory can thus be carried over to such a class of distributed structures. Ozaki and Ishii[2] applied such a transformation to obtain physical realizability conditions for such (i.e., uniform, lossless, and equal electrical lengths) transmission lines in cascade. The same results have been better formulated and extended by Riblet.[3] An interesting root-locus approach has been used by Seidel[4] to derive the realizability of insertion loss functions. Finally, Shih[5] has recently used the same idea to obtain some results in the time domain. All of these results are obviously directly applicable to cascades of *RC* transmission lines of equal electrical lengths again by a simple change of variable.

1.3 *Results*

An entirely new formulation in terms of the forward and backward waves in the component transmission lines of arbitrary lengths is developed in this paper. Such a formulation is then used to obtain several significant results. Specifically, these results include:

(*i*) A method of analysis which allows one to write down, by inspection, the system functions of the cascaded structures. The expressions for these functions are obtained explicitly in terms of the physical parameters (characteristic impedances, propagation constants, etc.) of the component lines.

(*ii*) Physical realizability conditions for system functions of cascaded transmission lines.

(*iii*) A synthesis method which is simple and appears to have the distinction of minimizing computational errors.

1.4 *Organization*

We begin with a statement of the problem in complete generality but we end up with restricting it to the case of interest here (i.e., cascades of uniform, lossless transmission lines). A summary of results for the equal length case follows. We then proceed to introduce our new formulation and discuss the lossless case in detail. The ideas developed for the lossless case will be extended to other structures in a subsequent paper.

II. STATEMENT OF PROBLEM

In its completely general form, a cascade of linear two-ports may be represented as in Fig. 1. Each component two-port may be characterized by any one of numerous relationships between the various signal parameters at the two ports. The most convenient one for a cascade structure relates all the signal parameters at one port to those at the other. The signal parameters that we shall use are the voltages and the currents at the several ports. Other parameters (such as forward and backward waves and many others) can also be used; but, these are not so con-

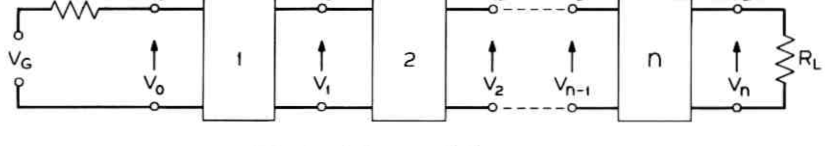


Fig. 1 — The cascaded structure.

venient. The conventions of positive directions for voltages and currents are also shown in the figure.

The input and output signal parameters for each two-port are then related by a transmission matrix for that two-port. Thus,

$$S_{k-1} = T_k S_k \tag{1}$$

where S_k is the signal vector whose elements are $\{V_k , I_k\}$ and T_k is the transmission matrix for the kth two-port. It follows that

$$S_o = T_1 T_2 \cdots T_n S_n = T S_n \tag{2}$$

and

$$T = T_1 T_2 \cdots T_n . \tag{3}$$

Equation (3) allows us to study the properties of the composite transmission matrix T in terms of those of the component matrices T_k . Our interest is in the methods of analysis and synthesis of such structures. These are carried out conveniently in terms of some scalar system function of the complex frequency variable $s = \sigma + j\omega$. The system functions that we shall be concerned with are the impedance function

$$Z_o(s) = \frac{V_0(s)}{I_0(s)} , \tag{4}$$

and the transmission (or insertion) loss function

$$\Theta(s) = \frac{V_G(s)}{2V_n(s)} \cdot \sqrt{\frac{R_L}{R_G}} \tag{5}$$

where V_G is the voltage of the source and appropriate resistive source and load terminations (R_G and R_L) are assumed. The above functions are simply related to the elements $t_{ij}(s)$ of the matrix T.

$$Z_o(s) = \frac{t_{11}R_L + t_{12}}{t_{21}R_L + t_{22}} \tag{6}$$

and

$$\Theta(s) = \frac{t_{11}R_L + t_{12} + t_{21}R_G R_L + t_{22}R_G}{2\sqrt{R_G R_L}} . \tag{7}$$

In this paper, our interest is limited primarily to those structures that are representable as cascades of uniform lossless transmission lines. The component two-ports are thus lossless transmission lines whose ends are the ports (Fig. 2). The component matrix T_k can now be obtained from the transmission line equations which, under zero initial

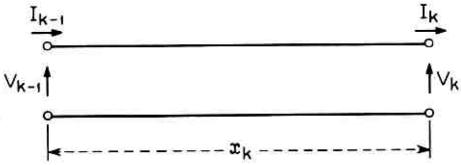


Fig. 2 — A single section of lossless transmission line.

conditions and Laplace transformation with respect to the time variable, are:[6]

$$\begin{bmatrix} \dot{V}(s) \\ \dot{I}(s) \end{bmatrix} = \begin{bmatrix} 0 & -sL_k \\ -sC_k & 0 \end{bmatrix} \begin{bmatrix} V(s) \\ I(s) \end{bmatrix} \tag{8}$$

where $\dot{V}$ and $\dot{I}$ are the derivatives with respect to the distance variable x. For the kth line, it follows that

$$\begin{bmatrix} V_{k-1}(s) \\ I_{k-1}(s) \end{bmatrix} = \begin{bmatrix} \cosh s\tau_k & R_k \sinh s\tau_k \\ R_k^{-1} \sinh s\tau_k & \cosh s\tau_k \end{bmatrix} \begin{bmatrix} V_k(s) \\ I_k(s) \end{bmatrix}, \tag{9}$$

where L_k and C_k are the inductance and capacitance per unit length of the line,

$$\tau_k = \sqrt{L_k C_k} x_k = \text{electrical length},$$
$$R_k = \sqrt{L_k / C_k} = \text{characteristic impedance},$$

and x_k = physical length of the kth line. We thus have

$$T_k = \begin{bmatrix} \cosh s\tau_k & R_k \sinh s\tau_k \\ R_k^{-1} \sinh s\tau_k & \cosh s\tau_k \end{bmatrix}. \tag{10}$$

We shall use (10) to derive most of our results.

III. EQUAL ELECTRICAL LENGTHS (LOSSLESS)

In this section, we briefly summarize the known results that have been obtained for the case of transmission lines of equal electrical lengths, i.e.,

$$\tau_k = \tau \qquad \text{for all } k. \tag{11}$$

Actually, τ has the dimension of time and it is the time of propagation in each line. It is commonly expressed as a fraction of the wavelength, hence it is called the electrical length.*

* In the sequel, it will simply be called "the length"; when physical length is meant, it will be so specified.

The transmission matrix T_k is now dependent only on the parameter R_k. Now, it is clear from (6) that any factor common to all t_{ij} cancels out in the expression for Z_o. If we make all matrices T_k rational in some variable, except for a scalar multiplier, then the function Z_o will also be rational. From (10) and (11), it is apparent that either the hyperbolic cosine or sine is the scalar multiplier if we use the transformation

$$p = \tanh s\tau,$$

or

$$= \coth s\tau. \tag{12}$$

The matrices T_k are then all rational in p (except, of course, for the scalar multipliers) and so, Z_o will also be a rational function in p.

It should be observed that (12) maps the real and imaginary axes into the real and imaginary axes, respectively, and the right half-plane into the right half-plane. It is this fact together with the rational Z_o that allows us to draw upon the theory of lumped networks. We summarize some of the important conclusions. First, we choose $p = \coth s\tau$ and observe that

$$Z_{k-1}(p) = \frac{pZ_k(p) + R_k}{R_k^{-1}Z_k(p) + p}, \qquad k = 1,2,\cdots,n \tag{13}$$

where $Z_k(p)$ is the impedance

$$Z_k(s) = \frac{V_k(s)}{I_k(s)} \tag{14}$$

under the above change of variables.

A basic theorem for the physical realizability of cascaded lossless equal length lines is as follows.

Theorem:[2,3] *The necessary and sufficient conditions that Z_o, a real rational function of p of degree n be the input impedance of cascaded lossless equal-length lines terminated in a resistor are: (i) Z_o is a positive-real function of p, and (ii) even part of Z_o has only the n-fold zeros at $p = \pm 1$.*

The necessity of condition (*i*) follows from (13) by observing that the real part of Z_{k-1} is non-negative for all values of p with non-negative real parts whenever the real part of Z_k is also non-negative for those values of p. By iteration of (13) the first condition is seen to follow. The second condition follows from the determinant of T_k which is $(p^2 - 1)$ if we neglect the scalar multiplier $\sinh s\tau_k$. From (3), the determinant of T is $(p^2 - 1)^n$, neglecting the scalar multiplier again. But the determinant of (10), is given by the difference of products of the even and the odd polynomials in the numerator and the denominator of the impedance function Z_o. This difference is also the numerator of the even part of Z_o and the second condition is seen to be necessary. Sufficiency of these conditions can be shown by an actual constructive synthesis procedure using the inverse relationship of (13), *viz.*,

$$Z_k = \frac{pZ_{k-1} - R_k}{p - R_k^{-1}Z_{k-1}}. \tag{15}$$

From (15), one can show that if Z_{k-1} satisfies the above conditions, then so does Z_k and it is of a lower degree. The process ultimately terminates yielding the load resistance.

Other results in the p-domain include explicit expressions for the coefficients of the input impedance in terms of the characteristic impedances of the lines and vice versa.[3] There is also some discussion on the realizability of the transmission loss functions.[3] These results follow from the basic theorem above.

An interesting departure from the above is the time domain investigation of the same structure.[5] No physical realizability conditions are available in the time domain; however, the synthesis procedure is conceptually quite simple. The system function used is the (impulse) reflection function in the time domain which takes the form of an infinite series of equally spaced impulses. The first impulse at $t = 0$ can only result from the first discontinuity thereby yielding R_1. The second impulse ($t = 2\tau$) results from the second discontinuity and yields R_2 (since we know R_1). The third impulse ($t = 4\tau$) results from the third discontinuity as well as multiple reflections encountering the first and second discontinuities. Since the only unknown in all these discontinuities is the third one, it is uniquely determined and yields R_3. The process continues and every new impulse determines the next characteristic impedance until all the junctions are specified. The rest of the impulses are then sums of the multiple reflections from all the junctions and the synthesis is complete.

The major drawback of the time domain approach is that there are no concise physical realizability conditions available.

IV. LOSSLESS CASE (GENERAL)

In this section, we consider the general case of lossless transmission lines of arbitrary lengths in cascade. The transformation (12) no longer reduces the system functions into rational functions. In fact, it is no longer possible to think in terms of rational functions. We must, therefore, abandon the previous approach and start fresh.

We begin with some physical observations. The structure is certainly passive and so the impedance function Z_o must be a positive-real function of s. The component two-ports as well as the cascade of them represent reciprocal structures and so the determinant of T_k as well as that of T must be unity. This follows from the reciprocal property in general and can be verified from (3) and (10) directly. The next observation stems from the time delay property of transmission lines. As mentioned earlier, the length of the transmission line τ_k represents in reality a time delay of τ_k seconds between the input and output signals for the kth line. The cascade structure, of course, distorts the signal but we can still speak of the time delay as the time interval between the start of the input signal and that of the output signal. This is the time delay that an impulse will undergo, *viz.*,

$$\tau = \sum_{k=1}^{n} \tau_k \, . \tag{16}$$

In this same cascade structure, however, each component line may be viewed as a delay line of length τ_k . To bring the parameter τ_k in prominence, we can look upon the line with its discontinuities at the two ends as a spatial resonator for an impulse. If we can make these elementary resonators τ_k explicitly apparent in the system functions, we would be able to identify the several lines. It is this fact that motivates the formulation that we shall pursue.

Let

$$z = e^{s} \tag{17}$$

so that

$$z^{\tau_k} = e^{s\tau_k}. \tag{18}$$

This maps the left half s-plane into the unit disc whose boundary $| z | = 1$ corresponds to the imaginary axis of the s-plane. Then

$$T_k = \tfrac{1}{2}[A_k^{+} z^{\tau_k} + A_k^{-} z^{-\tau_k}], \tag{19}$$

where

$$A_k^{\pm} = \begin{bmatrix} 1 & \pm R_k \\ \pm R_k^{-1} & 1 \end{bmatrix}. \tag{20}$$

Equation (19) expressed T_k directly in terms of the forward and backward wave delays $z^{-\tau_k}$ and $z^{+\tau_k}$. It would be more meaningful to express

(19) in terms of the delay terms, $z^{-\tau_k}$, and the terms, $z^{-2\tau_k}$, corresponding to the elementary resonator. However, we shall find positive exponents of z more convenient to use and when necessary it is always possible to revert to the negative exponents. Hence, we shall have occasion to use

$$T_k = \frac{1}{2z^{\tau_k}}\,[A_k^+ z^{2\tau_k} + A_k^-]. \tag{21}$$

V. LOSSLESS CASE — ANALYSIS

It is desirable in many cases to study the behavior of system functions for different values of physical parameters of the system. In such cases, it is necessary to bring out explicitly the dependence of these functions on the system parameters. We proceed to do so by first expressing T in terms of these parameters. From (3) and (19)

$$T = \prod_{k=1}^{n} \frac{1}{2}\,[A_k^+ z^{\tau_k} + A_k^- z^{-\tau_k}], \tag{22}$$

or

$$T = 2^{-n} \sum A_1^{u_1} A_2^{u_2} \cdots A_n^{u_n} z^{u_1\tau_1 + u_2\tau_2 + \cdots + u_n\tau_n} \tag{23}$$

$$\text{and } u_k = \pm 1 \text{ when a coefficient} \tag{24}$$

$$= \pm \text{ when a superscript.}$$

The summation above is over all possible combinations $(u_1, u_2, \cdots, u_n)$. Thus, there are 2^n terms in all. Each of these terms needs to be examined further to make (23) meaningful. First, however, let us observe that the matrix T as well as the functions Z_o and Θ can be all obtained very simply from

$$y = Tv, \tag{25}$$

where $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. For example,

$$Z_o = \frac{y_1}{y_2} \tag{26}$$

and

$$2\sqrt{R_G R_L}\,\Theta = y_1 + R_G y_2, \tag{27}$$

where, in (25),

$$v = \begin{pmatrix} R_L \\ 1 \end{pmatrix}. \tag{28}$$

The elements of matrix T are obtained by letting the vector v have elements $\{1,0\}$ and $\{0,1\}$. Our interest will, therefore, be in obtaining y in terms of v. It follows from (23) that we need merely obtain

$$v' = A_1^{u_1} A_2^{u_2} \cdots A_n^{u_n} v. \tag{29}$$

We show below that $A_k^{u_k}$ is singular and so v' is obtained by successive projections of a vector onto the appropriate eigenvector. Let

$$A_k^{u_k} = \begin{bmatrix} 1 & u_k R_k \\ u_k R_k^{-1} & 1 \end{bmatrix}. \tag{30}$$

It is obvious that the above matrix is singular ($u_k^2 = +1$). Its nonzero eigenvalue is at $\lambda = 2$ with the eigenvector

$$e_2^{u_k} = \begin{pmatrix} u_k R_k \\ 1 \end{pmatrix}; \tag{31}$$

its other eigenvector is

$$e_0^{u_k} = \begin{pmatrix} -u_k R_k \\ 1 \end{pmatrix}. \tag{32}$$

It should be clear that $A_k^{u_k}$ operating on any vector v results in two times the projection of v onto $e_2^{u_k}$. Or,

$$A_k^{u_k} v = u_k R_k^{-1}(v_1 + v_2 u_k R_k) e_2^{u_k}. \tag{33}$$

Then, from (29)

$$v' = e_2^{u_1} \left(v_2 + \frac{v_1}{u_n R_n} \right) \prod_{k=1}^{n-1} \left(1 + \frac{u_{k+1} R_{k+1}}{u_k R_k} \right). \tag{34}$$

Finally, we obtain

$$y = 2^{-n} \sum z^{u_1 \tau_1 + \cdots + u_n \tau_n} \left[\prod_{k=1}^{n-1} \left(1 + \frac{u_{k+1} R_{k+1}}{u_k R_k} \right) \right] \left(v_2 + \frac{v_1}{u_n R_n} \right) e_2^{u_1}, \tag{35}$$

where the summation is again over combinations $(u_1, u_2, \cdots, u_n)$. Equation (35) expresses the system functions as well as the composite matrix explicitly in terms of the system parameters τ_k and R_k. Further simplifications in (35) are possible for special situations. However, the important thing to be emphasized here is that we have an explicit expression in scalar form for the elements y_1 and y_2 and therefore for all system functions of interest. For computational purposes, (35) can be expressed in terms of hyperbolic cosine and sine terms. For discussing physical realizability, it would be more convenient to eliminate all negative exponents of z in (35). This is accomplished by using (21) or, equivalently, by considering $z^{\tau}y$ since the highest negative exponent in (35) is τ. The impedance function Z_o will be now a ratio of functions involving only positive exponents of z.

$$Z_o(z) = \frac{z^{\tau}y_1}{z^{\tau}y_2}. \tag{36}$$

VI. LOSSLESS CASE — PHYSICAL REALIZABILITY

The basic results will be derived for the realizability of the impedance function $Z_o(z)$. It is then easy to carry over the results to determine the realizability of other system functions. Let Z_o be expressed in the form

$$Z_o(z) = \frac{N(z)}{D(z)} = \frac{\sum_{k=0}^{m} a_k z^{2i_k}}{\sum_{k=0}^{m} b_k z^{2i_k}}, \tag{37}$$

where N and D are finite sums as shown and have no common factors, i_k are nonnegative and increasing with k. The coefficients a_k and b_k are real and both are not zero for any k.

The necessary conditions that must be satisfied have been mentioned before (see Section IV):

(*i*) Z_o must be a positive-real function of s, or

$$\text{Re } Z_o(z) \geqq 0 \qquad \text{for} \qquad |z| \geqq 1.$$

(*ii*) Determinant of T is one. Since we are considering $(z^{\tau}T)$

$$\text{det. } (z^{\tau}T) = z^{2\tau} = z^{2i_m}.$$

(see (23), (36), and (37)).

The second condition must be somehow expressed in terms of N and D. To do this, observe that except for a constant positive multiplier, $z^{\tau}T$ is a product of matrices of the type

$$T_k' = \begin{bmatrix} (z^{2\tau_k} + 1) & R_k(z^{2\tau_k} - 1) \\ R_k^{-1}(z^{2\tau_k} - 1) & (z^{2\tau_k} + 1) \end{bmatrix} = \begin{bmatrix} f_1^k(z) & f_2^k(z) \\ g_2^k(z) & g_1^k(z) \end{bmatrix} \tag{38}$$

where for all k

$$\begin{aligned} f_1(z) &= z^{2\tau_k} f_1(1/z); \qquad & g_1(z) &= z^{2\tau_k} g_1(1/z) \\ f_2(z) &= -z^{2\tau_k} f_2(1/z); \qquad & g_2(z) &= -z^{2\tau_k} g_2(1/z) \end{aligned} \tag{39}$$

and

$$\text{det. } T_k' = f_1 g_1 - f_2 g_2 . \tag{40}$$

Observe that the product of T_k' will yield

$$T' = \begin{bmatrix} F_1(z) & F_2(z) \\ G_2(z) & G_1(z) \end{bmatrix} \tag{41}$$

where F_1, F_2, G_1, and G_2 satisfy the same type of relations as f_1, f_2, g_1, and g_2, respectively, *viz.*,

$$\begin{aligned} F_1(z) &= z^{2\tau} F_1(1/z) \\ F_2(z) &= -z^{2\tau} F_2(1/z), \qquad \text{etc.} \end{aligned} \tag{42}$$

Also, if $c > 0$ is a constant,

$$z^{2\tau} = \text{det. } (z^{\tau} T) = c \text{ det. } (T') = c(F_1 G_1 - F_2 G_2). \tag{43}$$

Then, if

$$N = N_1 + N_2 \tag{44}$$

and

$$D = D_1 + D_2 ,$$

where

$$\begin{aligned} N_1(z) &= z^{2i_m} N_1(1/z); \qquad & D_1(z) &= z^{2i_m} D_1(1/z) \\ N_2(z) &= -z^{2i_m} N_2(1/z); \qquad & D_2(z) &= -z^{2i_m} D_2(1/z), \end{aligned} \tag{45}$$

we can express, using (41),

$$Z_o = \frac{N_1 + N_2}{D_2 + D_1} = \frac{F_1 R_L + F_2}{G_2 R_L + G_1} . \tag{46}$$

The condition (2) now can be expressed using (43) and (46) as

$$N_1 D_1 - N_2 D_2 = c z^{2i_m} \tag{47}$$

where c is again a positive constant.

It is further possible to simplify the statement of the necessary conditions. Z_o is a positive-real function for $|z| \geqq 1$. Consequently, it is also an analytic function for all $|z| \geqq 1$, hence one need verify the non-negative property of Z_o only on the boundary $|z| = 1$.* On the unit

* For a justification of all such statements, see the Appendix.

circle,

$$\operatorname{Re} Z_o \,|\,_{|z|=1} = \tfrac{1}{2}[Z_o(z) + Z_o(1/z)]_{|z|=1}\,. \tag{48}$$

Define

$$EvZ_o(z) = \tfrac{1}{2}[Z_o(z) + Z_o(1/z)];$$

then, using (44)–(46),

$$\begin{aligned} EvZ_o(z) &= \frac{1}{2}\left[\frac{N_1(z) + N_2(z)}{D_2(z) + D_1(z)} + \frac{N_1(z) - N_2(z)}{-D_2(z) + D_1(z)}\right] \\ &= \left[\frac{\{N_1(z)D_1(z) - N_2(z)D_2(z)\}}{D(z)D(1/z)z^{2i_m}}\right]. \end{aligned} \tag{49}$$

Substituting (47) in (49), we have

$$EvZ_o(z) = \frac{c}{D(z)D(1/z)}, \qquad (c \geqq 0). \tag{50}$$

Observe that the real part of Z_o is always positive since c is positive. It is zero only if c is zero and this can happen only if $R_L = 0$ or ∞. The two necessary conditions are thus equivalent to:

(*i*) $D(z)$ is Hurwitz-type, i.e., all its zeros lie in the interior of the unit circle.

(*ii*) $EvZ_o = c[D(z)D(1/z)]^{-1}; \qquad c \geqq 0.$

These alternative conditions are easier to check. In any case, we now state and prove the physical realizability conditions in the following theorem.

Theorem: The necessary and sufficient conditions that $Z_o(z)$ be an impedance function of a resistively terminated cascade of lossless, uniform transmission lines are:

(*i*) *$Z_o(z)$ is a positive real function for $|z| \geqq 1$.*

(*ii*) $N_1D_1 - N_2D_2 = cz^{2i_m}, \qquad (c \geqq 0, \quad i_m > 0)$.*

Proof: The necessity of the conditions has already been shown. The sufficiency will be shown by a constructive method of realization. In fact, we shall show that given a Z_o satisfying these conditions, it represents the impedance of a transmission line terminated in an impedance Z_1 satisfying the same conditions and of lower order. Z_o and Z_1 are re-

* A lossless structure will result if $c = 0$.

lated by

$$Z_o = \frac{(z^{2\tau_1} + 1)Z_1 + R_1(z^{2\tau_1} - 1)}{(z^{2\tau_1} - 1)Z_1R_1^{-1} + (z^{2\tau_1} + 1)} \tag{51}$$

or

$$Z_1 = -\frac{(z^{2\tau_1} + 1)Z_o - R_1(z^{2\tau_1} - 1)}{(z^{2\tau_1} - 1)Z_oR_1^{-1} + (z^{2\tau_1} + 1)}. \tag{52}$$

To show that Z_1 is *p-r* and of lower order for some positive R_1 and τ_1 we first observe, from (44) through (46), that

$$Z_o(\infty) = \frac{N_1(0) - N_2(0)}{D_1(0) - D_2(0)} = -\frac{N_2(0)}{D_1(0)},$$

since by condition (2),

$$\frac{N_1(0)}{N_2(0)} = \frac{D_2(0)}{D_1(0)};$$

and

$$\begin{aligned} Z_o(0) &= \frac{N_1(0) + N_2(0)}{D_1(0) + D_2(0)} = \frac{N_2(0)}{D_1(0)} \\ &= -Z_o(\infty). \end{aligned} \tag{53}$$

Next, we observe from (52) that

$$\rho_1' = \frac{\dfrac{Z_1}{R_1} - 1}{\dfrac{Z_1}{R_1} + 1} = \frac{\dfrac{Z_o}{R_1} - 1}{\dfrac{Z_o}{R_1} + 1} z^{2\tau_1} = \rho_0' z^{2\tau}. \tag{54}$$

In the above, both ρ_1' and ρ_0' are reflection coefficients. It is obvious that if ρ_1' is analytic for $|z| \geqq 1$ and bounded by one on the unit circle, than Z_1 is *p-r*. It is also true that if ρ_1' is of lower order than ρ_0' then Z_1' is of lower order than Z_o'. We, therefore, let $R_1 = Z_o(\infty)$ and note that the highest exponents of z in the numerator and the denominator of ρ_0' are $2i_{m-1}$ and $2i_m$, respectively. The denominator of ρ_0' has no constant term and has the lowest exponent of $2i_1$. If we now choose τ_1 equal to the lesser of i_1 and $(i_m - i_{m-1})$, it is assured that ρ_1' is analytic for $|z| \geqq 1$ and well behaved at infinity. This follows from the analyticity of ρ_0' and the cancellation of $z^{2\tau_1}$. It is also clear from (54) that for $|z| = 1$, $|\rho_1'| \leqq |\rho_0'|$. But since Z_0 is *p-r*, $|\rho_0'| \leqq 1$, hence $|\rho_1'|$ is bounded by one. We thus have, Z_1 is *p-r* if Z_o is *p-r* and the order of Z_1 is, $2(i_m - \tau_1)$, when the order of Z_o is $2i_m$.

Next, if we express

$$Z_1 = \frac{N_1' + N_2'}{D_1' + D_2'} \tag{55}$$

where N_1', N_2', D_1', and D_2' are defined in a similar manner as N_1, N_2, D_1, and D_2 in (45), except that

$$N_1'(z) = z^{2(i_m - \tau_1)} N_1'(1/z), \qquad \text{etc.} \tag{56}$$

It then follows from (51) and (55) that

$$\begin{aligned} N_1(z) &= (z^{2\tau_1} + 1)N_1'(z) + R_1(z^{2\tau_1} - 1)\, D_2'(z) \\ N_2(z) &= (z^{2\tau_1} + 1)N_2'(z) + R_1(z^{2\tau_1} - 1)\, D_1'(z) \\ D_1(z) &= (z^{2\tau_1} + 1)D_1'(z) + R_1^{-1}(z^{2\tau_1} - 1)N_2'(z) \\ D_2(z) &= (z^{2\tau_1} + 1)\, D_2'(z) + R_1^{-1}(z^{2\tau_1} - 1)N_1'(z). \end{aligned} \tag{57}$$

From condition (2),

$$\begin{aligned} cz^{2i_m} &= N_1D_1 - N_2D_2 \\ &= 4z^{2\tau_1}(N_1'D_1' - N_2'D_2') \qquad \text{(from (57))}. \end{aligned}$$

Thus,

$$N_1'D_1' - N_2'D_2' = c'z^{2(i_m - \tau_1)}, \qquad c' \geqq 0$$

and the second condition is satisfied. This proves the basic theorem.

VII. LOSSLESS CASE — SYNTHESIS

It is indeed possible to synthesize the cascaded structure using (52) to (54) as discussed in the previous section. We present here an algorithm for synthesis which is much more straightforward. In our discussion here, we shall tacitly assume that the conditions of the realizability theorem are satisfied. Given a $Z_o(z)$ satisfying the realizability conditions there exist R_1 and τ_1 such that it is the impedance function of a transmission line of length τ_1 and characteristic impedance R_1 terminated in a realizable impedance function Z_1 of order lower than that of Z_o. Let

$$Z_o = \frac{y_1}{y_2}$$

and

$$Z_1 = \frac{y_1'}{y_2'};$$

then if y is a vector with components $\{y_1, y_2\}$, v_k are some vectors, and

$$y' = \sum_{k=0} v_k z^{2i_k},$$

we have

$$\begin{aligned} y &= \tfrac{1}{2}(A_1{}^{+}z^{2\tau_1} + A_1{}^{-})y' \\ &= \tfrac{1}{2}(A_1{}^{+}y')z^{2\tau_1} + \tfrac{1}{2}(A_1{}^{-}y'). \end{aligned}$$

The ratio of elements in the first term on the right is $+R_1$ and for the second term it is $(-R_1)$. The lowest exponent of z in the first term is $2\tau_1$. Finally, y' is obtained by removing the multiplier $z^{2\tau_1}$ in the first term and adding the terms together since

$$\tfrac{1}{2}(A^{+} + A^{-}) = I,$$

the identity matrix. We thus have a unique algorithm provided we have a nondegenerate structure, i.e., the sum of the lengths of any subset of the lines is not equal to the sum of the lengths of any other subset. This assures us that there are no exponents equal in the two terms above. We shall now specify the algorithm.

Given an impedance function.

$$Z_o = \frac{\sum_{k=0}^{m} a_k z^{2i_k}}{\sum_{k=0}^{m} b_k z^{2i_k}};$$

(i) Separate like and unlike signs of the coefficients a_k, b_k

$$Z_o = \frac{\sum a_l z^{2i_l} + \sum a_u z^{2i_u}}{\sum b_l z^{2i_l} + \sum b_u z^{2i_u}},$$

$$\frac{a_l}{b_l} > 0; \qquad \frac{a_u}{b_u} < 0.$$

(ii) Identify $a_l/b_l = R_1$ and the lowest $i_l = \tau_1$.
(iii) Obtain

$$Z_1 = \frac{\sum a_l z^{2(i_l-\tau_1)} + \sum a_u z^{2i_u}}{\sum b_l z^{2(i_l-\tau_1)} + \sum b_u z^{2i_u}}.$$

The algorithm is repeated until step (iii) leads to a constant representing the terminating resistor R_L. It must be observed that this algorithm is valid for nondegenerate structures only (i.e., $a_l/b_l = -(a_u/b_u)$ for all l and u).

For degenerate structures, the first step in the algorithm has to be modified. It is known, of course, from our discussion of (53) that

$$\frac{a_m}{b_m} = -\frac{a_0}{b_0} = R_1 .$$

So, if for any k, $(a_k/b_k) \neq \pm R_1$, then we must split a_k and b_k such that

$$a_k = a_{kl} + a_{ku}$$

$$b_k = b_{kl} + b_{ku}$$

and

$$\frac{a_{kl}}{b_{kl}} = -\frac{a_{ku}}{b_{ku}} = R_1 ,$$

so that

$$a_{kl} = R_1 b_{kl} = \tfrac{1}{2}(a_k + R_1 b_k)$$

$$a_{ku} = -R_1 b_{ku} = \tfrac{1}{2}(a_k - R_1 b_k).$$

Using the above, we obtain the modified algorithm:

(*i*) Identify

$$\frac{a_m}{b_m} = R_1 .$$

(*ii*) Decompose

$$Z_o = \frac{\sum a_{kl} z^{2i_k} + \sum a_{ku} z^{2i_k}}{\sum b_{kl} z^{2i_k} + \sum b_{ku} z^{2i_k}} .$$

(*iii*) Identify the lowest i_k with nonzero $a_{kl} = \tau_1$.

(*iv*) Obtain

$$Z_1 = \frac{\sum a_{kl} z^{2(i_k - \tau_1)} + \sum a_{ku} z^{2i_k}}{\sum b_{kl} z^{2(i_k - \tau_1)} + \sum b_{ku} z^{2i_k}} .$$

The synthesis method presented here minimizes algebraic operations on the coefficients a_k and b_k, hence it is computationally advantageous.

VIII. CONCLUSION

We have presented a formulation which allows us to investigate structures involving lossless transmission lines of arbitrary electrical lengths. An analysis method is then developed which explicitly expresses the system functions in terms of the physical parameters of the system.

A basic theorem specifying the physical realizability conditions for such structures has been presented together with a computationally simple method of synthesis of impedance functions satisfying these conditions. The significant characteristic of the results presented so far is the simplicity of the algorithms involved both for analysis as well as synthesis. These algorithms allow one to proceed by inspection in simple problems and are most suitable for computer studies when the problems are more complex.

Extensions of the theory to more general transmission lines and lumped structures have been carried out. These results as well as design approaches to the cascade structures and questions of testing conditions, approximations, etc., will be discussed elsewhere.

APPENDIX

Maximum Modulus Theorem and Transcendental Functions

Throughout the text, the maximum modulus theorem[7] has been applied to functions which have either essential singularities or are not single-valued in the domain concerned. Some justification for the validity of the theorem for such functions is in order. The theorem has been used to imply that the unit bound on the reflection coefficient (or the positive reality of the impedance function) on the imaginary axis of the s-plane is sufficient to ensure the same throughout the semi-infinite right half s-plane. Consider the reflection function

$$\rho(s) = \frac{\sum_{n=0}^{m} a_n e^{2s i_n}}{\sum_{n=0}^{m} b_n e^{2s i_n}}, \qquad b_m \neq 0$$

where i_n are nonnegative and increasing with n. The above function is, of course, assumed to be analytic in the right half-plane. The function $\rho(s)$ is a meromorphic function with infinite singularities, hence the point at infinity is an essential singularity. This makes it difficult to apply the maximum modulus theorem to the entire right half-plane. The transformation $z = e^s$ eliminates the essential singularity at infinity but makes $\rho(z)$ multi-valued since i_n are not necessarily integers. If the i_n are indeed integers, then $\rho(z)$ is single-valued and the theorem can be applied. If the i_n are not integers, they can be approximated arbitrarily closely by rational numbers (dense in the field of real numbers) and the transformation $z^u = e^s$, where $ui_n =$ integer, will yield a single-valued function to which the theorem can be applied. This discussion should suffice to justify the use of the maximum modulus theorem for our purposes. In fact, the theorem can be applied to the function in its original s-domain or under any suitable transformation.

REFERENCES

1. Richards, P. I., Resistor-Transmission-Line Circuits, Proc. IRE, *34*, **1948**, pp. **317**.
2. Ozaki, H. and Ishii, J., Synthesis of Transmission-Line Networks and the Design of UHF Filters, IRE Trans., *CT-2*, **1955**, pp. **325**.
3. Riblet, H. J., General Synthesis of Quarter-Wave Impedance Transformers, IRE Trans., *MTT-5*, **1957**, pp. **36**.
4. Seidel, H., Synthesis of a Class of Microwave Filters, IRE Trans., *MTT-5*, **1957**, p. **107**.
5. Shih, S. T., Synthesis of Optical Filters by Transmission-Line Analogs, M. S. Thesis, MIT, **1965**.
6. Weber, E., *Linear Transient Analysis*, John Wiley and Sons, Inc., New York, N. Y., **1956**, Vol. II, Chapter 6.
7. Titchmarsh, E. C., *The Theory of Functions*, Oxford University Press, London, **1939**, Chapter V.

Electron Phase Contrast Images of Molecular Detail

By R. D. HEIDENREICH

(Manuscript received March 14, 1966)

Electron phase contrast images with a resolution of at least 2 Å have been obtained using a modified commercial electron microscope. A "phase column" approximation for interpreting such images is briefly discussed and applied to images of graphite, evaporated carbon, and a synthetic polypeptide. Hexagonal features about 5 Å in diameter are attributed to the graphite unit cell imaged by the six first-order prism plane reflections. The image so produced is a next-nearest neighbor representation.

The steady improvement in resolving power of commercial electron microscopes over the past few years has re-awakened considerable interest in the possibilities of directly imaging details of molecular structure. In particular, the phase contrast mechanism based on the Abbé theory of image formation expressed in terms of the Kirchoff diffraction integral has been re-examined using numerical computation methods not in wide use when Scherzer[1] discussed phase image formation. Several theoretical papers[2,3,4,5] dealing with phase contrast images of atom positions have indicated that it should be possible to experimentally obtain such images under the right conditions. The computations all assume a monolayer specimen in the object plane or effectively a single atom approach. This idealized specimen is difficult to realize experimentally and for that reason this brief account is concerned with problems and some results with actual three-dimensional preparations.

High resolution Fourier or "lattice" images of near perfect three-dimensional crystals[6,7] about one-fourth extinction distance thick have become familiar in the last two to three years with image detail exhibited down to 1.8 Å. These have been obtained with crystals for which the total elastic cross section considerably exceeds the inelastic. Tilted illumination has been used which effectively reduces the spherical aberration to zero in one direction. Since it appears that the greatest potential value of high resolution microscopy lies in molecular biology, the subjects of concern here are carbon and high polymers for which the elastic and inelastic cross sections are about the same and the phase contrast situation is, thus, not so favorable.

There are two aspects to high-resolution microscopy: first, an objective lens and instrument capable of point-to-point resolution in the range of interatomic distances and, second, preparation of specimens and interpretation of phase contrast images to obtain intelligible information from the object; i.e., the ability to use the resolving power.

The micrographs displayed here were taken with a modified* Siemens Elmiskop I having improved stability, reduced ac hum, and a focal length shortened from 2.8 to 1.9 mm. The result is a resolving power (point-to-point) with axial illumination of at least 2 Å. In the present configuration the instrument is operated at 80 KV and double condenser with a 200 μ condenser aperture to improve the transverse coherence for phase contrast. Images were recorded photographically at 214,000X and always in focal sequences at 35 Å focal steps.

In the coherent phase-amplitude approximation-to-phase contrast, the intensity $| \Psi |^2$ at a point (x^i,y^i) in the image plane is

$$| \Psi(x^i,y^i) |^2 \cong M^{-2} [1 + | S |^2 + 2S^{\text{real}}] \tag{1}$$

with M the magnification and S^{real} the real part of the phase-amplitude or Kirchoff imaging integral[4,5] over the back focal plane. S is the integral of the product of several terms,† one of which is sin χ with χ the phase relative to the unscattered axial wave.

$$\chi \equiv \chi_{\text{sph.}} + \chi_{\text{astig.}} + \frac{k}{2} \Delta f \beta^2 . \tag{2}$$

Here, $\chi_{\text{sph.}} \approx -(\pi/2\lambda)C_s\beta^4$ is the spherical aberration phase, β the scattering angle, and Δf the defocus from the precise Gaussian image condition. Since $C_s = Cf$, where f is the focal length of the objective lens, a reduction in f reduces the phase distortion due to spherical aberration. The phase shift $\pi/2$ due to scattering is removed from (2) and included in S. In order that the imaging integral have a useful magnitude for an interplanar spacing $d = | g |^{-1}$, the specimen thickness t must be such that

$$\sqrt{\lambda t/2} \ll d \tag{3}$$

to hold down the destructive phase summation over the specimen thick-

* The modifications were volunteered and carried out by Mr. H. Armbruster of Siemens-America. The author is grateful to Mr. Armbruster for his skill, effort and time in improving the instrument.

† A discussion of the various factors involved with three-dimensional objects will be given in a full publication.

ness. Generally, the thinner the object, the better the phase contrast resolution. Objects 50 Å thick or less are desirable which poses problems in preparation and mounting techniques.

An instructive approach to the interpretation of phase contrast detail is to employ the column approximation[8] used so successfully in diffraction contrast. The difference will be in the fact that the transmitted and diffracted beams from the column are recombined at the image plane and that the column will have definite dimensions and symmetry. These dimensions and symmetry are determined by the reciprocal lattice vectors **g** accepted by the objective lens and not subject to undue phase distortion. Since the angle β in (2) for a Bragg reflection is $\beta = \lambda \mid g \mid$, the imaging integral S will have appreciable magnitude in the column for values of Δf and $\mid g \mid$ that bring (2) near $\pi/2$. The column of thickness t will thus have a prismatic cross section with dimensions and symmetry determined by the **g** vectors optimized by defocus Δf for a given spherical aberration coefficient.

For the case of a thin sheet of graphite normal to the optic axis of the objective lens, the electron diffraction pattern consists only of reflections (hko). Of these, only the six prism plane reflections at 2.13 Å are used by the objective lens to produce an image without undue phase distortion. Shorter spacings are presently seriously "garbled" by both spherical aberration and uncorrected astigmatism. Micrographs taken with and without a 100 μ objective aperture, which passes the prism plane reflections, are very similar. The result is a phase contrast column of thickness t and hexagonal cross section about 5 Å in diameter as depicted in Fig. 1. The relation between the minimum crystallographic unit cell and the column cross section is evident in Fig. 1. Since no information on the shorter interatomic distances reaches the image plane, the cross section is that for a "next-nearest neighbor cell". The nearest neighbor cell would be essentially a benzene ring shown by the dash lines in Fig. 1. It is noted that the information available to the image plane recognizes only four atoms in the cross section of the column whereas there are ten carbon atoms in the structure.

The column approach just discussed is highly useful in interpreting the micrograph of graphite in Fig. 2. The size of the hexagonal "cells" in Fig. 2 is about 5 Å as expected on the basis of Fig. 1. Inelastic scattering and some damage to the thin graphite due to cleaving and mounting act to degrade the delineation of the image of the cells.

The high "noise" level familiar in evaporated carbon substrates becomes understandable from this point of view. A typical high resolution micrograph of a thin carbon substrate is displayed in Fig. 3. In the circled

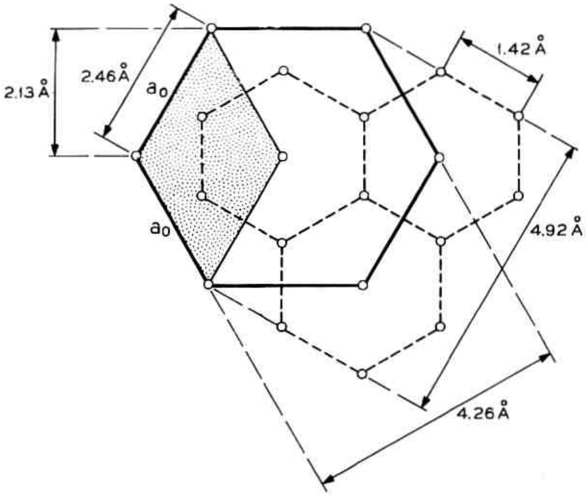


Fig. 1 — Hexagonal array of atoms in a single graphite layer. The unit cell of side a_0 = 2.46 Å (c = 6.7 Å) is shown cross-hatched. The hexagonal cell shown by the heavy lines is the one defined by the six prism plane reflections at 2.13 Å or the next-nearest neighbor cell.

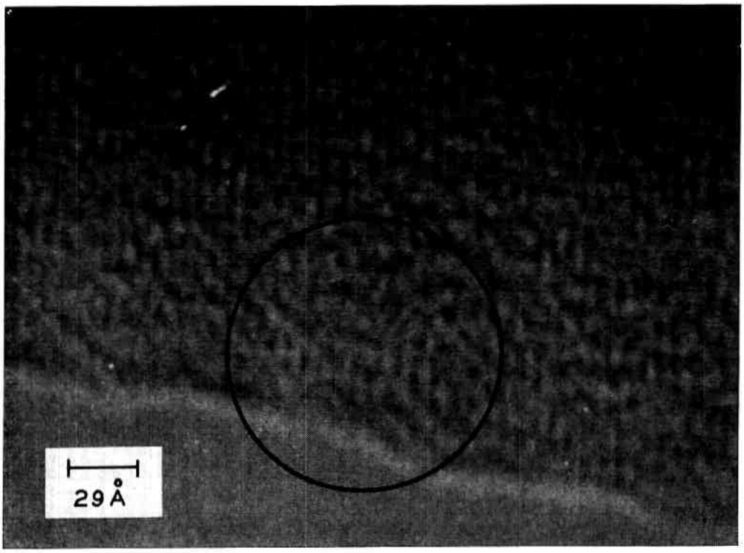


Fig. 2 — Phase contrast micrograph of a cleaved graphite sheet using the prism plane reflection and showing the hexagonal cells of Fig. 1 about 5 Å in diameter. The defocus is about 100 Å to the focal length side. (80 *KV*, 200 μ condenser aperture. No objective aperture. Electronic magnification 214,000×.)

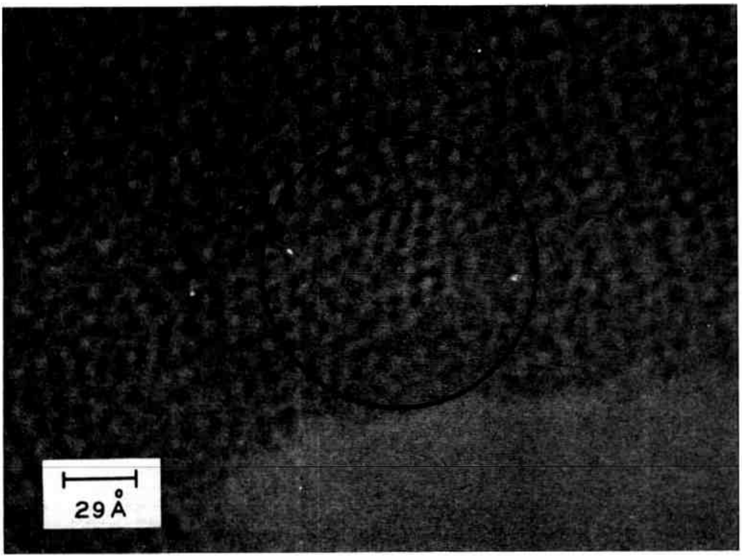


Fig. 3 — Micrograph of an evaporated carbon substrate displaying the hexagonal cells in the circled area where the *c*-axis is normal to the sheet. The crystallite size in this region is around 20 Å. Neighboring regions are at different orientations.

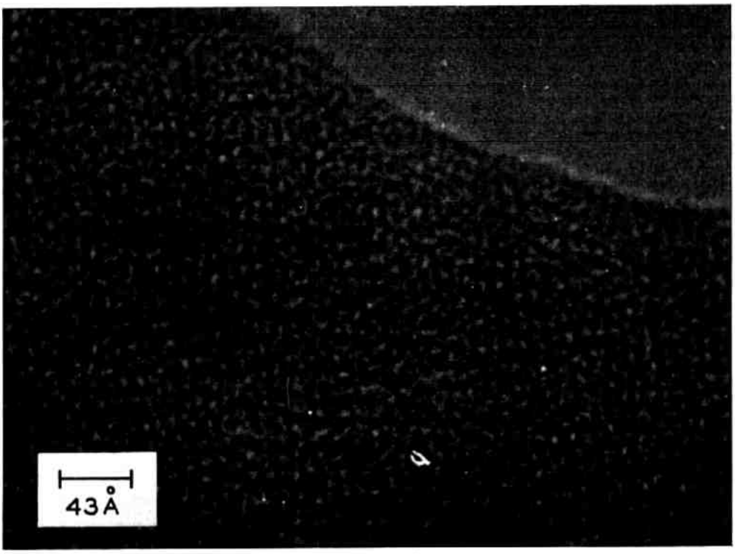


Fig. 4 — Micrograph of a thin film (≈60 Å) of the synthetic polypeptide poly-γ-benzyl-L-glutamate at lower magnification showing the extent of ordered structure seen by phase contrast. Hexagonal cells such as those of Fig. 3 are found scattered through the image.

region are several hexagonal cells about 5 Å in diameter taken to be due to a small graphite crystallite only a few cells in extent oriented with the c-axis normal to the film. Neighboring areas are composed of other crystallites unsuitably oriented to produce the prism plane reflections. So called amorphous carbon is well approximated by randomly oriented graphite crystallites only a few cells in extent. The structure seen in carbon films is thus actually "graphite noise".

All the polymer films and filaments examined under these conditions show phase contrast structure. Although the use of cold surfaces about the object holder reduces contamination they do not completely remove it. In addition, most polymers suffer radiation damage by 80 KV electrons which leads to the appearance of a diffuse diffraction ring at about 2.1 Å. Consequently, there can be some question as to the origin of hexagonal cells seen in the polymers. There is sufficient detail in such images, however, differing from that seen in Fig. 3 that the images cannot be regarded as just due to carbon. The micrograph in Fig. 4 is a phase image of a thin film of poly-γ-benzyl-L-glutamate (a synthetic polypeptide) at about half magnification of the preceding figures. The ordered structure in the image is evident, but at present the details have not been explained. The α-helix is about 12 Å in diameter so that a single chain should be easily seen if it were isolated. The chains must, therefore, be packed together in Fig. 4 and may be cross-linked which greatly increases the difficulty in interpretation.

From a number of micrographs such as Fig. 4, it appears that if phase contrast high resolution microscopy is to be useful for polymers and biological molecules, techniques of preparing specimens to give isolated chains must be developed. The phase image of an isolated chain should be quite amenable to interpretation using the column concept.

REFERENCES

1. Scherzer, O., J. Appl. Phys. *20,* 1949, p. 20.
2. Hoppe, W., Naturwissenshaften, *48,* 1961, p. 736.
3. Lenz, F., Optik, *21,* 1965, p. 489.
4. Heidenreich, R. D. and Hamming, R. W., Numerical Evaluations of Electron Image Phase Contrast, B. S. T. J., *44,* 1965, p. 207.
5. Eisenhandler, C. and Siegel, B. M., J. Appl. Phys., Feb., 1966.
6. Komoda, T., J. Electron Microscopy, *14,* 1965, p. 128.
7. Komoda, T., Optik, *21,* 1964, p. 93.
8. Hirsch, P. B., Howie, A., Nicholson, R. B., Pashley, D. W., and Whelan, M. J., *Electron Microscopy of Thin Crystals,* Butterworths, London, 1965, p. 157.

Contributors to This Issue

L. E. Franks, B.S., 1952, Oregon State University; M.S., 1953, and Ph.D., 1957, Stanford University; Bell Telephone Laboratories, 1958—. Mr. Franks has been engaged in communication and network theory studies related to data transmission systems. He was a visiting lecturer in signal theory at Columbia University for the spring term of 1965. Member, IEEE, Sigma Xi.

R. D. Heidenreich, B.S., 1938, M.S., 1940, Case Institute of Technology; Bell Telephone Laboratories, 1945—. His work has been chiefly in the areas of electron microscopy and electron diffraction. He developed the thin metal section methods for transmission electron microscopy now widely used for studying defects in solids. His early application of electron methods to semiconductors resulted in chemical polishing techniques and long surface lifetime treatments for germanium. He has conducted extensive joint research programs on magnetic materials which have correlated structure with magnetic anistrophy in both hard and soft permanent magnets. His more recent theoretical studies concerning elastic and inelastic scattering of electrons has led to his present interest in high-resolution electron imaging aimed toward resolving atomic configurations. He is at present on leave of absence to teach electron microscopy during the spring semester at Georgia Institute of Technology. Member, AAAS; Fellow, American Physical Society; Past President, Electron Microscope Society of America.

R. A. Kimber, B.S.M.E. 1959, University of Illinois; M.S.M.E. 1961, New York University; Bell Telephone Laboratories, 1959—. Mr. Kimber has done development and the analytical work on clip-type connectors and is currently responsible for the development of a new coin chute for the 1A1 Coin Telephone. Member, Tau Beta Pi, Pi Tau Sigma, Sigma Tau; Registered Professional Engineer, Indiana.

B. K. Kinariwala, B.S., 1951, Benares University (India); M.S., 1954, and Ph.D., 1957, University of California; Bell Telephone Laboratories, 1957—. Mr. Kinariwala was first engaged in research in circuit theory involving, in particular, active and time-varying networks. More recently, he has been concerned with problems in communication systems. Member, IEEE, Sigma Xi.

John G. Skinner, H.N.C., M.E., 1948 and H.N.C., Physics, 1950, Northampton Polytechnic, London, England; M.S., 1958, and Ph.D., 1962, Oregon State University; Bell Telephone Laboratories, 1961—. Mr. Skinner has been engaged in the study of solid-state lasers and optical deflection schemes. Member, Sigma Xi, Phi Kappa Phi, Optical Society of America.

R. R. Stokes, B.S.M.E. 1953, Clemson University; Bell Telephone Laboratories, 1953—. Mr. Stokes has been engaged in work in the Automatic Reporting Telephone (ART) and Card Dialer. He is presently Supervisor in the Public Telephone Department, responsible for custom engineered coin telephone products. Member, Tau Beta Pi, Phi Kappa Phi.

S. Y. Tong, B.S.E.E., 1955, Taiwan University; M.S.E.E., 1961, University of Vermont; Bell Telephone Laboratories 1964—. Mr. Tong is completing his Ph.D. at Princeton University. He has been concerned with the problems in coding theory and in the field of fault detection and diagnosis in digital computers. Member, IEEE; Associate member, Sigma Xi.

Aaron D. Wyner, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, Ph.D., 1963, Columbia University; Bell Telephone Laboratories, 1963—. Mr. Wyner has been engaged in research in various aspects of information theory. He is also Adjunct Assistant Professor of Electrical Engineering at Columbia University. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

B.S.T.J. BRIEFS

Multicolor Holographic Image Reconstruction with White-Light Illumination

By L. H. LIN and K. S. PENNINGTON and G. W. STROKE* and A. E. LABEYRIE*

Color images have been obtained by wavefront reconstruction from a reflection volume hologram illuminated with ordinary white light. The hologram was recorded with coherent light at two wavelengths, 6328 Å and 4880 Å, from helium-neon and argon-ion lasers, respectively. Fig. 1 shows the white-light reconstructed image from such a hologram; the original subject was a color transparency. The hologram was formed in Kodak 649F emulsion.

A simple method of multicolor holography has previously been reported.[1] This method was based upon the formation of volume holograms which reconstructed by Bragg reflection from the planes formed in the emulsion. The wavefronts were reconstructed by illuminating the hologram with the same laser light used in recording and were observed on transmission through the hologram plate. With beam angles used to give transmission, the Kodak 649F emulsion was not thick enough to form holograms having the angular and spectral selectivities needed for good white-light reconstruction. A simple method for obtaining reflection volume holograms was recently described.[2] It showed that high-quality reconstructions could be obtained in a single color by reflection of white light from the hologram when, in the recording, the reference beam and the subject beam interfered at very large angles (160°–180°). Reflection holograms of two- and three-dimensional objects form an extension of basic ideas and work by Denisyuk[3] in his generalization of Lippmann color photography[4] and Gabor holograph.[5]

The ability to reconstruct multicolor holograms with white-light illumination adds a degree of flexibility to holography; we have now demonstrated the simplicity of obtaining this result. We have recorded reflection holograms both by "projection" and in diffused light, in a single color and in multiple colors. Fig. 2 illustrates one of the arrangements used to record the multicolor hologram. To insure minimum shrinkage of the emulsion in processing the hologram, we omitted the fixing of the emulsion as suggested by Ives.[6] Any white light source rang-

* University of Michigan.

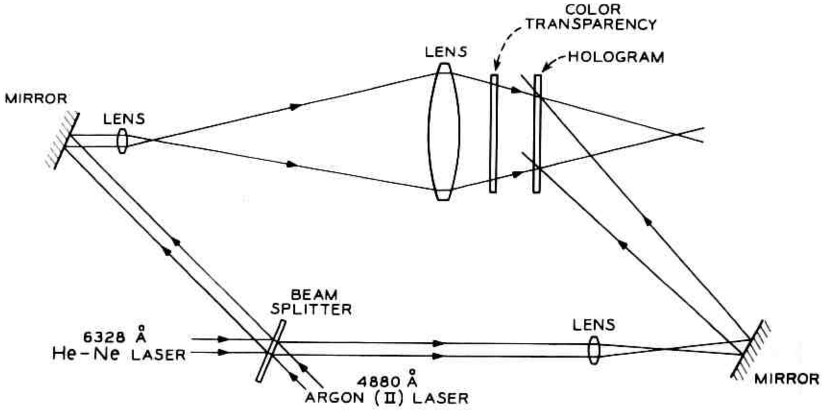


Fig. 2 — Arrangement for "projection" hologram formation.

ing from flashlight to sunlight can be used to obtain reconstructions as shown in Fig. 1. Particularly brilliant multicolor reconstructions were obtained when the light illuminating the subject was focused some distance behind the hologram plane. A similar result was reported in Ref. 2.

One of us (G. W. Stroke) wishes to thank Professor D. Gabor for most fruitful conversations and encouragement with this work. He also wishes to acknowledge the generous support by the National Science Foundation of the part of the work carried out by him with his students.

REFERENCES

1. Pennington, K. S. and Lin, L. H., Applied Physics Letters, *7*, 1965, pp. 56–7.
2. Stroke, G. W. and Labeyrie, A. E., Physics Letters, *20*, 1966, pp. 367–369.
3. Denisyuk, Yu. N., Soviet Physics-Doklady, *7*, 1962, p. 543.
4. Lippmann, G., J. de Physique, *3*, 1894, p. 97.
5. Gabor, D., Nature, *161*, 1948, p. 777.
6. Ives, H. E., Astrophys. J., *27*, 1908, p. 325.

Fig. 1 (Opposite page) — Early photograph of multicolor holographic image reconstructed with white light.

Note added in proof: The color of this photograph was shifted toward the blue. More faithful recording of the reconstructed wavefront has been obtained with a better choice of color film and angle of the reconstructing illumination.