THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 51 April 1972 Number 4

Conditional Vertical Subsampling– A Technique to Assist in the Coding of Television Signals

By R. F. W. PEASE

(Manuscript received July 23, 1971)

In an interlaced scan television system, the vertical sampling rate of an image can be halved by sampling every other field. Picture elements in the missing fields are replaced in the display by both temporal and vertical interpolation, but the resulting pictures show some visible defects. This paper describes how these defects can be eliminated at the extra cost of fully sampling in selected areas of the picture. For a typical Picturephone® *scene with active movement the selected areas make up about 6 percent of the picture elements in the unsampled field. The technique can be combined with a wide variety of interframe-coding techniques. In one particular example in which the television signal is specified as clusters of frame-to-frame differences, the cost of specifying "active" frames (14,000 significant frame differences per frame) is reduced from 68,000 bits to 42,500 bits. This corresponds to a reduction in bit rate from 2 Mbits sec^{-1} to 1.3 Mbits sec^{-1}.*

I. INTRODUCTION

The technique of reducing the horizontal-sampling frequency ("subsampling") in the moving parts of the television image has been pre-

viously described.[1] It was found that the frequency could be halved without visible degradation for most object speeds. For slow speeds the degradation was visible but not objectionable. Combining this technique with that of conditional replenishment[2] has proved particularly effective,[3] because in periods of fast movement, conditional replenishment by itself becomes uneconomic because so many picture elements (pels) change significantly from frame to frame.

An obvious question to ask is whether a similar advantage results from subsampling vertically.

In an interlaced television scan format, the idea of halving the vertical-sampling frequency can be confusing because we must distinguish between fields and frames. A diagram showing the vertical position of lines for successive fields is shown in Fig. 1. One method of halving the vertical-sampling frequency is to sample every second line in each field and to replace the unsampled lines by interpolating the values of vertically adjacent elements in the same field [e.g., an unsampled line with coordinates $y = 2$ and $t = 2$ is replaced by an average of lines with y, t coordinates (0, 2) and (4, 2)]. This has been tried and the degradation is subjectively objectionable.

The second method is to sample alternate fields so that in stationary pictures the vertical-sampling frequency is halved. The unsampled fields are replaced by an average of the four nearest neighbors in Fig. 1 (e.g., an unsampled line with coordinates $y = 2$ and $t = 2$ is replaced by an average of lines with y, t coordinates 1 , 1; 1 , 3; 3 , 1; and 3 , 3). With this method the resulting pictures are subjectively satisfactory except for fast moving contrasty edges which appear blurred and somewhat jerky, and in dark areas with little horizontal but much vertical detail;[4] in these areas an aliasing pattern is sometimes visible.

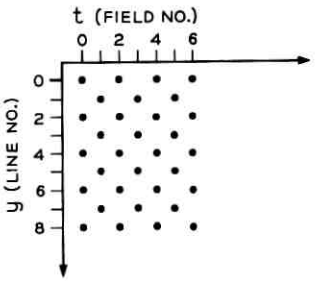

Fig. 1—Temporal position (or Field No., t) and vertical position (or Line No., y) of lines in a 2:1 interlace scan.

Henceforth, we will use the term "vertical subsampling" to refer to this second method.

If we vertically subsample over most of the moving regions except where there are visible defects, the entire picture should be subjectively satisfactory. We call this technique, conditional vertical subsampling (CVSS). The question is: How much extra channel capacity is needed to bring about the required improvement? This paper describes some experiments designed to answer this question.

Because the extra information is generated at an irregular rate, the use of conditional vertical subsampling requires a buffer memory and hence is probably most useful in conjunction with a buffered coder. In this paper we have in mind the eventual use of conditional vertical subsampling in a buffered interframe coder similar to those described in Refs. 2 and 3.

II. EXPERIMENTAL ARRANGEMENT

The basic apparatus is shown in Fig. 2. The output of the television camera (a silicon diode array vidicon) is sampled at 2.02 MHz and is digitized to 8-bit accuracy. The scan format is similar to that used in the *Picturephone* service in that there are 271 lines per frame and 60 fields per second. Each frame is made up of two interlaced fields.

Consider first only the odd-numbered fields. The coded version of a previous odd field appears at the output of field delay 1 and is compared, picture element by picture element, with the digitized input to determine whether the stored value is an adequate representation of the current value. For any pel, if the difference between the input signal and the stored signal exceeds a value ($T1$) of four levels out of 255, then $t_1 = 1$ and S1 is switched to the "one" position so that the value of the input replaces the previous value in the field delay. This new value is circulated twice in the field delay (because S1 is held at 0 during even fields) and after exactly one frame time is compared with the new input signal. Usually, in the absence of movement, the comparison is good enough so that no further updating need take place (i.e., no new information need be sent to a hypothetical receiver). Thus, each odd field is conditionally replenished as described in Ref. 2.

When the replenished version of line 3 , 3 (i.e., line No. 3, field No. 3) emerges first from switch S1, the replenished version of line 1 , 3 is appearing at the output of the line delay. Similarly, the replenished version of line 3 , 1 is appearing at the output of field delay 1 and the line 1 , 1 is appearing at the output of the second half-line delay. Thus,

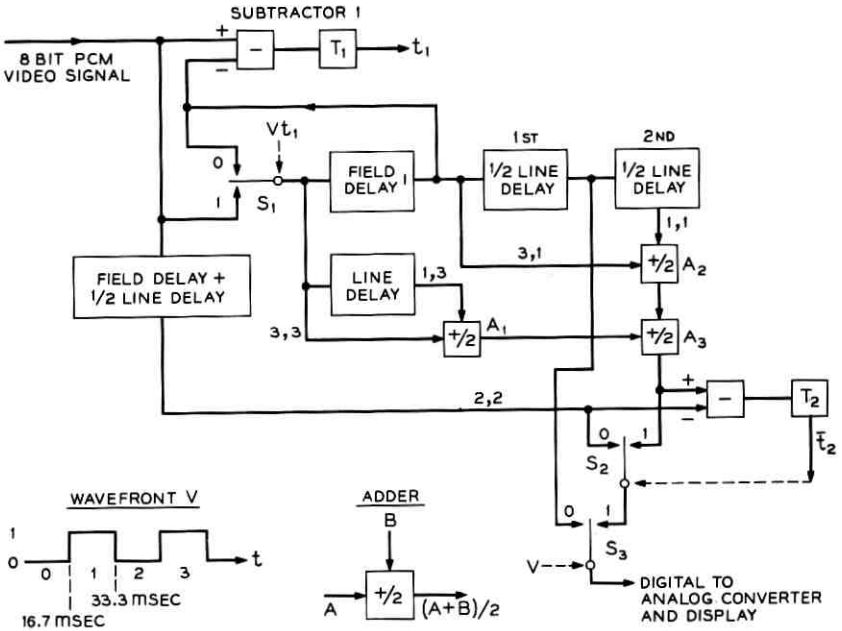

Fig. 2—Basic diagram of apparatus used to evaluate conditional vertical subsampling. Odd fields are coded by conditional (frame) replenishment using the loop formed by switch S1 and field delay 1; the output for these fields appears at 0 on switch S3. The output for even fields appears at 1 on switch S3 and is derived either from the vertically subsampled signal at the output of adder A3 or from the original output delayed by 1 field and 1/2 line.

lines 3 , 3; 1 , 3; 3 , 1; and 1 , 1 are simultaneously available and so, using adders A1, A2, and A3, we can form the spatially and temporally interpolated value, which as described in the previous section, is displayed instead of line 2, 2 when there is subsampling.

To maintain the correct temporal and spatial sequence of displayed fields, the replenished version of line 3 , 3 is not displayed until one field time plus one-half line time after first appearing at switch S1. Thus, line 3, 3 and other lines in odd fields are taken from the output of the first one-half line delay, and if we wish to vertically subsample continuously, the output is taken from adder A3 for even fields (i.e., S2 is kept in the 1 position and S3 is switched by waveform V).

However, we want to investigate how much the picture quality is improved by replacing the interpolated value with the original signal in selected parts of the even fields. To do this in the proper time sequence, the original signal is delayed by one field time plus one-half line time

and then compared with the interpolated value in subtractor 2. When the magnitude of the difference equals or exceeds the threshold $T2$, then $t_2 = 1$ and the original is displayed. Otherwise the interpolated value is displayed. In a practical coder the position and amplitude of those parts in the original signal used to replace the interpolated value must be coded and transmitted to the receiver. It is the cost of transmitting this extra information which must be balanced against the resulting improvement in picture quality.

The output of switch S3, which corresponds to the output of the hypothetical receiver, is converted to an analog signal and displayed on a television monitor with a 5-1/2 inch by 5-inch raster and a polarizing faceplate. The picture was viewed at 36 inches in a room with average illumination for an office (70 foot-candles).

Three series of experiments were carried out. In the first series, the subjective effect of varying the threshold $T2$ was investigated. In the second series, we measured the frequency of occurrence of picture elements for which the difference between the interpolated value and the delayed signal equals or exceeds $T2$ (henceforth we shall refer to such events as VSS differences). In the third series, we investigated the subjective and numerical effects of grouping the VSS differences into clusters so that a separate address word need not be used for each VSS difference. For most experiments requiring numerical results, the scene was the swinging model head shown in Fig. 3. The period of the swing was 2.7 seconds and the amplitude corresponded to 44 picture elements. The maximum speed of the head corresponded to 3-1/2 pels per frame interval. From time to time, some experiments were also carried out with live subjects or with very contrasty material.

III. RESULTS AND DISCUSSIONS

3.1 *Subjective Effects of Varying the Threshold T2*

The results of the two extremes is already known; for $T2 = 1$, we have the original picture and for $T2 = 255$, we have the vertically subsampled picture with the aforementioned defects. With a threshold of $T2 = 4$ (out of 255 levels), close scrutiny (15 inches) of the hair of the model when stationary revealed a just detectable difference from the original only if the original was compared instantaneously; otherwise there was no difference between the two pictures. With $T2 = 8$, there was a slight loss in vertical detail in the hair of the model which again was only noticeable at the normal viewing distance by switching instantaneously to the original. In some areas containing a lot of dark

Fig. 3—Swinging model head scene. The head is swung with a peak amplitude shown by the distance between adjacent vertical pencils.

vertical detail but little horizontal detail, the low contrast aliasing pattern could be made out. For $T2 = 12$, there was a more visible loss of vertical detail of stationary pictures and fast movement (4 pels per frame interval and above) of contrasty edges (64 levels per pel) showed some smearing or occasional raggedness. For $T2 = 16$, the above effects were more pronounced. As the threshold was increased beyond 16, the above effects became progressively more serious. For $T2 = 32$, the picture quality was virtually the same as for $T = 255$.

3.2 *Frequency of Vertical Subsampling Differences For Even Fields*

In even fields we counted the number of VSS differences for various values of $T2$. The numbers were recorded on a digital printer with a sampling frequency of about 5 Hz; the scene was the swinging model head. Simultaneous counts were also made of the frame differences exceeding a threshold of 4 (out of 255 levels) occurring in the odd fields. This count served as a check for the repeatability of the scene and also gave a measure of the amount of activity in the scene.

Some of the results are shown in Fig. 4. The dotted lines indicate the frequency of frame differences and show the characteristic periodic

pattern as the head swings. The variability in the depth of the trough is probably due to the relatively low sampling frequency. The maximum activity generates more than 7000 significant frame differences in the odd field which, in comparison with other *Picturephone* scenes is considered "active."[3] The corresponding curves for the vertical subsampling differences show the same form of activity but are much less variable. For instance, for $T2 = 8$ the variation is from 1100 per field at the end of the swing to 1900 per field at the bottom of the swing. This is understandable as those vertical subsampling differences which arise from the spatial interpolation tend to remain unchanged as the movement in the picture decreases.

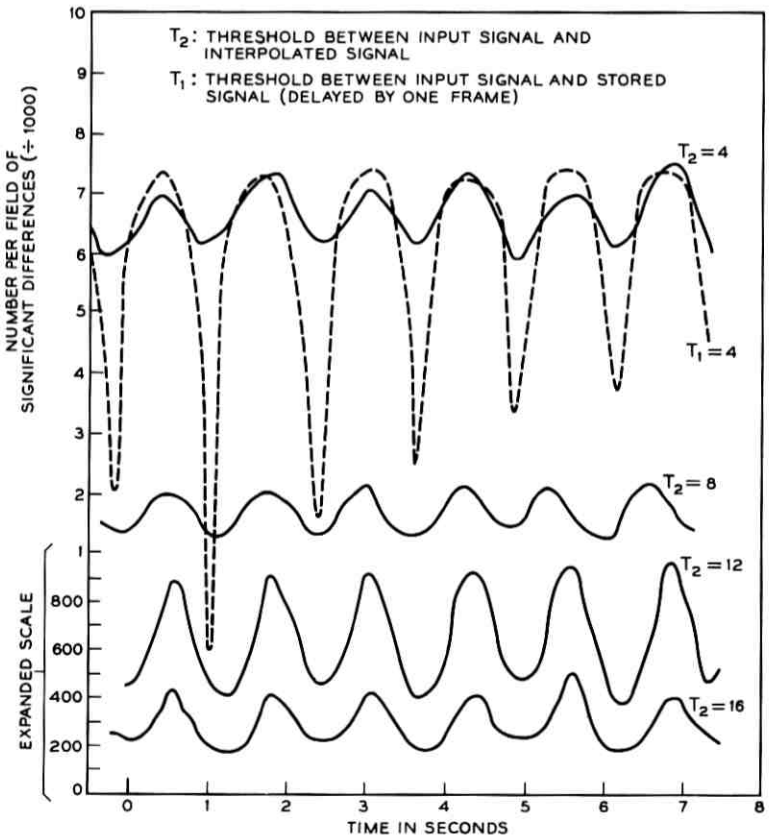

Fig. 4—Frequency of significant differences as a function of the scene (swinging model head).

Note the big decrease in counts of VSS differences as the threshold increases. For $T2 = 8$, which for a scene with movement gives a picture virtually indistinguishable from the fully sampled picture, there are only about 2000 significant VSS differences compared with more than 7000 significant frame differences. For $T2 = 16$, which still gives a substantial improvement in picture quality over that with $T2 = 255$, the frequency of significant VSS differences is less than 500 which is negligible compared with the frequency of significant frame differences.

If the threshold $T1$ (defining a significant frame difference) is raised to 8, then the picture quality is visibly degraded due to the "dirty window effect."[2] One other effect is to reduce the frequency of significant frame differences from about 7000 per field to about 5000 per field when the model head is at the bottom of the swing (Fig. 5). A third effect of increasing $T1$ from 4 to 8 is to increase the counts of significant VSS differences when $T2 = 8$ from about 2000 to 2600 per field. This increase can be seen by comparing the relevant graphs in Figs. 4 and 5. The increase is not noticeable for values of $T2$ greater than 10. This third effect can be easily explained by the increased difference between the respective odd fields and the original signal being added to the difference introduced by interpolation. Note however, that as both thresholds increase from 4 to 8, the counts of VSS differences decreased more than the corresponding counts of significant frame differences in spite of the fact that the subjective degradation of increasing $T1$ is more severe than that due to increasing $T2$.

3.3 *Cost of Transmission*

One way to transmit the extra information would be to use 8 bits to specify the horizontal position and another 8 bits to specify the revised amplitudes of each picture element showing a significant VSS difference. However, this is probably wasteful because: (*i*) the significant VSS differences tend to occur in clusters so that usually one address word can serve several VSS differences; the cost of this method is that either the length of the cluster or the end of the cluster must also be sent to the receiver. (*ii*) 8 bits are probably unnecessary for the amplitude information as the extra information can probably be adequately specified as a relatively coarsely quantized difference signal.

3.3.1 *Effect of Cluster Coding*

Direct observation of the spatial distribution of significant VSS differences (Fig. 6) shows that they, like frame differences, tend to occur in clusters and so cluster coding is probably advantageous.

To accentuate this clustering and to reduce further the number of bits required for addressing, it is usually advantageous to run two closely separated clusters into one larger cluster.[3] This is particularly important because the VSS differences due to temporal interpolation near a horizontally moving edge occur as two separate clusters (with differences of opposite sign) on either side of the edge but at the edge itself the differences are usually less than significant. This effect can be seen in Fig. 6 around the edges of the cheek of the model head; the reader may wish to verify exactly how this comes about by drawing out the video signal corresponding to a moving step on a target of a storage type television camera for three successive fields or by referring to the Appendix.

It also follows that isolated significant VSS differences are relatively

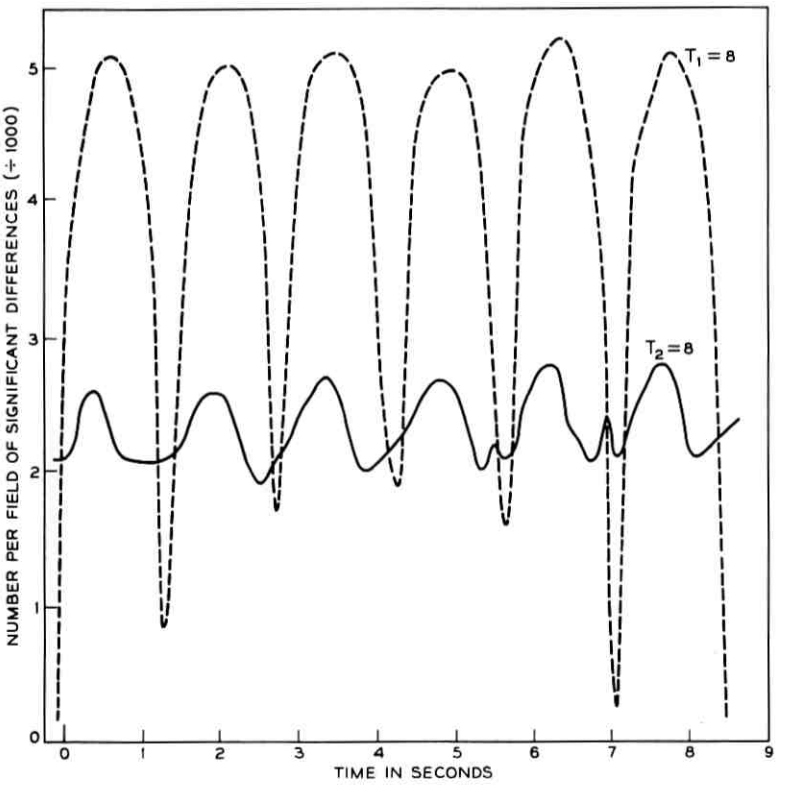

Fig. 5—Frequency of significant differences as a function of time. Scene is a swinging model head.

Fig. 6—Flags representing VSS differences for $T2 = 8$ while the head is swinging.

expensive to transmit, and they should not be transmitted if such rejection does not cause subjective degradation of the picture.[3] We investigated this effect by leaving uncorrected single VSS differences with no neighboring VSS differences within 2 pels horizontally. The subjective effect was negligible except for values of $T2$ of 12 and above, when occasionally the loss in vertical detail could be made out or a moving edge appeared somewhat more ragged. To control switch S2 in Fig. 2, the subtractor and threshold circuit were followed by logic to reject isolated VSS errors and to run together clusters of VSS differences separated by one or two pels. We also rejected VSS differences in the first and last lines of even fields because vertical interpolation here will be very frequent but rejecting these values is subjectively negligible. With these restrictions, we measured the frequency of clusters and of pels contained in clusters. Some of the results are summarized in Table I and indicate that the average length of clusters remained at about 5 irrespective of picture motion.

For comparison, equivalent data are shown for frame differences (using the same logic in recording the counts) and show that when the

scene is active the cluster length is longer for frame differences. These data correspond very well with equivalent data reported for scenes with live subjects.[3]

3.3.2 *The Effect of Quantizing the Signal*

Amplitude information in television signals is usually most economically specified by quantizing a prediction error. For example, the amplitude of the previously scanned pel is often used as a prediction of the current pel; the difference is then quantized and coded for transmission. This technique could be used here but we decided instead to quantize the prediction error generated by the interpolated value and the input. Such errors tend to be smaller than either element-to-element or frame-to-frame differences. In the Appendix, we show how this comes about in one particular case.

In Fig. 7 we have replotted some of the data of Fig. 4 and some extra data to show the frequency distribution of amplitudes of VSS differences for the swinging head. The curve is more peaked than comparable curves of element differences or frame differences and shows very few differences of amplitude greater then 10 percent of the peak amplitude. Thus quantizing the VSS difference directly is one attractive approach and should cost no more than 3 or 4 bits per pel if fixed length codes are used and perhaps much less if variable lengths are used.

As a first step, we divided the magnitude of the difference amplitude scale up into seven classes as shown in Table II and then assigned weights as shown to each class. The pictures that resulted from correcting the VSS errors with these quantized signals showed little difference from those in which the VSS errors are corrected with an 8-bit signal when the scene of the swinging head and the threshold 8 was used. Thus, a fixed code length of 3 bits would be sufficient to describe the amplitude information for each pel and still allow one word to denote the end of a cluster. However, when the scene contained a very

TABLE I—FREQUENCY OF FRAME DIFFERENCES, VSS DIFFERENCES, AND CLUSTERS FOR DIFFERENT THRESHOLDS FOR THE SWINGING HEAD SCENE (FIG. 2)

	End of Swing				Bottom of Swing				
Type of Difference	Frame	VSS	VSS	VSS	Frame	VSS	VSS	VSS	
Threshold	$T1 = 4$	8	12	16	4	8	12	16	
No. of Differences	123	750	350	230	7000	1400	650	400	Per Field
No. of Clusters	47	125	60	35	440	260	140	80	Per Field

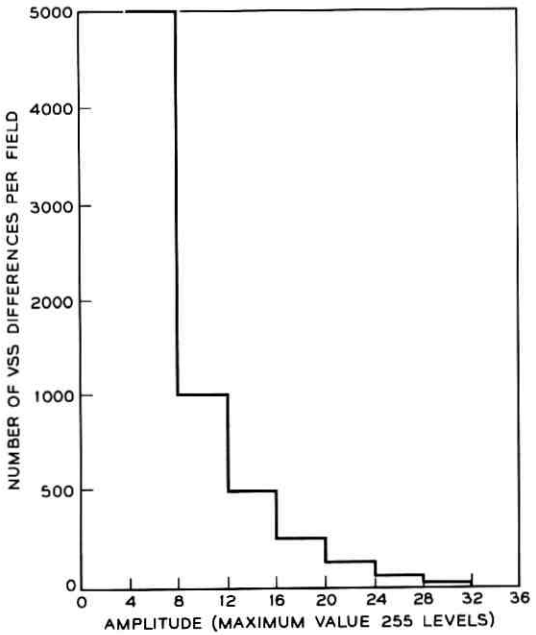

Fig. 7—Histogram of number of VSS differences as a function of amplitude values of 4 and above.

contrasty (50 percent of peak amplitude) edge moving at 4 pels per frame interval or faster, there was visible degradation of the edge. Therefore, a 15-level quantizer with the decision levels and weights shown in Table III was used so that a 4-bit vocabulary would suffice for the amplitude information and one word would be left to denote the end of a cluster. Resulting pictures were indistinguishable from those using an 8-bit correction signal.

If for each cluster we allow 4 bits per pel, 8 bits for the horizontal address, and 4 bits to denote the end of run, then we can plot the data which was summarized in Table I to show the total number of bits per field (excluding synchronizing bits) needed to specify the extra

TABLE II—7-LEVEL QUANTIZING SCALE

Decision Level	0	±8	±15	±25	} out of 255
Representative Weight	0	±10	±19	±28	}

TABLE III—15-LEVEL QUANTIZING SCALE

Decision Level	0	±3	±6	±10	±15	±23	±33	±44	out of 255
Representative Weight	0	±4	±8	±12	±18	±28	±38	±50	

information for correcting significant VSS differences (Fig. 8). For $T2 = 8$, which gives very satisfactory correction of the VSS differences, and for a frame with 14,000 significant frame differences, the even field can be sent at a cost of only 8500 bits. For $T2 = 12$ the cost comes down to 4500 bits and for $T2 = 16$ there is a further drop to 1600 bits.

For comparison, the cost, allowing 4 bits per frame difference and 12 bits per cluster of transmitting the frame differences, is also shown.

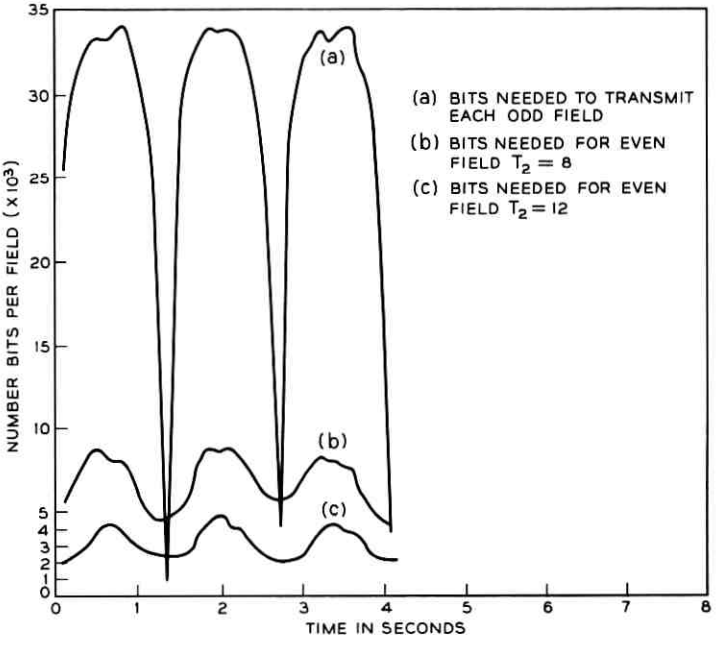

Fig. 8—Number of bits required to transmit each odd field by conditional replenishment and each even field by conditional vertical subsampling. For two values of $T2$. Scene: Swinging model head. In each field the pels requiring transmission are grouped in clusters by first rejecting isolated changes and then running together clusters separated by 1 or 2 pels. Twelve bits are assigned to each cluster and 4 bits to each pel containing a cluster. Synchronizing bits have been excluded.

Thus even though the frame differences occur in longer clusters in active frames, the cost of transmitting the odd field is much greater than the cost of transmitting the even field.

IV. CONCLUSIONS

For the scene used we can now answer the original question of comparing the advantages of horizontal subsampling and vertical subsampling by comparing the total cost of transmitting one frame (excluding synchronizing bits and "forced replenishment"[2,3] bits). For vertical subsampling we need, for an active frame, 34,000 bits for the odd field and 8,500 bits in the even to give a subjectively pleasing picture; this gives a total of 42,500 bits. When the moving parts of the picture are horizontally subsampled, the number of bits used to specify amplitude is halved but the number required for addressing is unchanged. For two fields, therefore, there are 28,000 bits for the amplitude information and 12,000 bits for the address information to give a total of 40,000 bits. Thus, the numerical advantage of the two techniques is very similar.

It should be pointed out that in the experiments described the vertical subsampling was applied to both the moving and stationary part of the picture. With a buffered interframe coder, the main problem is created by the moving parts of the picture. If the vertical subsampling is applied only to the moving part then the number of bits needed during the even fields is reduced still further. However, as can be seen from Fig. 6, there are relatively few VSS differences in the background and so in this case the savings would be marginal. The maximum savings so gained can be estimated by assuming that the only VSS differences that need correcting are those due to movement. The number of such differences can be estimated by assuming that they account for the difference between the number of VSS differences occurring for stationary scenes and for moving scenes. Thus in our experiment and for $T2 = 8$, the number of bits in the even field is approximately halved, and the total number of bits required to code the frame is reduced from 42,500 to 38,000.

One advantage of the technique of conditional vertical subsampling is that it can be combined with a wide variety of interframe-coding techniques, because the odd fields are not affected. One exception is that it is no longer straightforward to express amplitude information in the odd fields as field-to-field differences.[5] Otherwise the amplitude information in the odd fields can be expressed as element-to-element differences,[6] frame differences,[3] or even two-dimensional spatial differences.[7] How well the techniques of horizontal and vertical sub-

sampling can be simultaneously applied remains to be investigated. Horizontal subsampling in the odd field will certainly introduce extra VSS differences, but the numerical effect of these can probably be minimized by horizontally subsampling in the even fields as well.

V. ACKNOWLEDGMENT

I would like to acknowledge the many stimulating discussions with J. C. Candy, D. J. Connor, C. C. Cutler, L. H. Enloe, B. G. Haskell, J. O. Limb, and F. W. Mounts. The technical assistance of W. G. Scholes is also greatly appreciated.

APPENDIX

Vertical Subsampling Differences Arising From the Horizontal Movement of a Vertical Edge

Consider a vertical edge, consisting of a black to white transition of 128 levels, being moved horizontally across the field of view of a television camera at a speed of p pels per frame interval when referred to the target of the camera. Assume that the video-signal level for each pel is a direct measure of the quantity of light which has fallen on the appropriate area of the target for 1/30 second just prior to being scanned. If the target is stationary, then a plot of signal level as a function of horizontal position is shown by the line ABCD in Fig. 9. If immediately after pel 0 is scanned, the edge moves horizontally to the right at 8 pels per frame interval, then in the next frame the signal level will be a measure of the amount of time during the intervening 1/30th of a second that each point on the target received light from the bright side of the edge. For uniform motion therefore, the line ABED represents the signal level in the second frame. In the third frame the signal level is represented by the line AHD which is simply a translation of ABED by 8 pels. The signal level for the adjacent lines in the interlaced field is shown by the line AFGD and is simply a translation of ABED by 4 pels (there being no vertical difference). When vertically subsampling however, we substitute for the interlaced field, an interpolated signal ABD and the VSS differences are represented by the vertical distances between lines BD and BFGD. Note that at the halfway point this difference is zero and on either side of the edge we have relatively large differences with a maximum value of 32 levels. This is in agreement with the "flags" shown in Fig. 6, which occur in pairs of clusters straddling the edge of the moving cheeks.

We can also use Fig. 9 to compare certain differences in level which

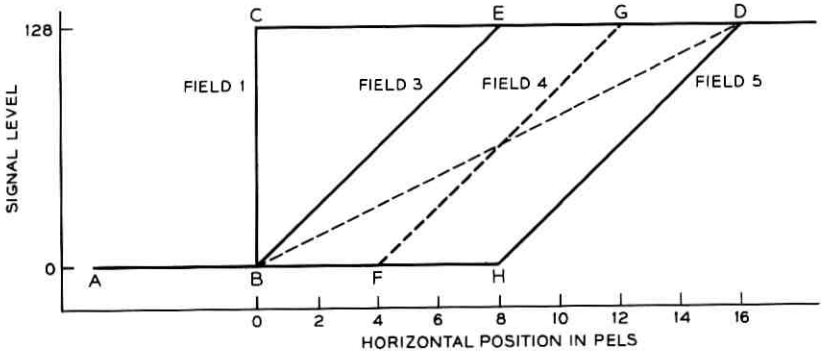

Fig. 9—Signal levels as a function of horizontal position for different fields.

may be quantized or coded for transmission. First of all, element-to-element differences: For a stationary position we have a maximum step height of 128 levels which decreases as the speed increases.

Frame-to-frame differences for a steadily moving edge are represented by the vertical distances between lines ABED and lines ABHD. The differences here have values up to 128 levels (a more accurate analysis shows that this level is only approached for values of $p \gg 1$).

Field-to-field differences are represented by the vertical distance between ABED and AFGD and can take values up to 64 levels.

Thus, VSS differences have, for the model, the smallest maximum value and would seem to be most economical to quantize and code.

REFERENCES

1. Limb, J. O., and Pease, R. F. W., "Exchange of Spatial and Temporal Resolution in Television," B.S.T.J., *50*, No. 1 (January 1971), pp. 191–201.
2. Mounts, F. W., "Video Encoding System with Conditional Picture Element Replenishment," B.S.T.J., *48*, No. 7 (September 1969), pp. 2545–2553.
3. Candy, J. C., Franke, M. A., Haskell, B. G., and Mounts, F. W., "Transmitting Television as Clusters of Frame-to-Frame Differences," B.S.T.J., *50*, No. 6 (July/August 1971), pp. 1889–1917.
4. Limb, J. O., and Pease, R. F. W., "A Simple Interframe Coder for Video Telephony," B.S.T.J., *50*, No. 1 (July/August 1971), pp. 1877–1888.
5. Pease, R. F. W., and Scholes, W. G., "Field Difference Quantization," unpublished work.
6. Limb, J. O., and Mounts, F. W., "Digital Differential Quantizer for Television," B.S.T.J., *48*, No. 7 (September 1969), pp. 2583–2599.
7. Connor, D. J., Pease, R. F. W., and Scholes, W. G., "Television Coding Using Two-Dimensional Spatial Prediction," B.S.T.J., *50*, No. 3 (March 1971), pp. 1049–1061.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

The MacWilliams Identities for Nonlinear Codes

By Mrs. F. J. MACWILLIAMS, N. J. A. SLOANE, and J.-M. GOETHALS

(Manuscript received December 13, 1971)

In recent years a number of nonlinear codes have been discovered which have better error-correcting capabilities than any known linear codes. However, very little is known about the properties of such codes. In this paper we study the most basic property, the weight enumerator. The weight of a codeword is the number of its nonzero components; the weight enumerator gives the number of codewords of each weight, and is fundamental for obtaining the error probability when the code is used for error-correction on a noisy channel. In 1963 one of us showed that the weight enumerator of a linear code is related in a simple way to that of the dual code (Jessie MacWilliams, "A Theorem on the Distribution of Weights in a Systematic Code," Bell System Technical Journal, 42, No. 1 (January 1963), pp. 79–94). In the present paper, which is a sequel, we show that the same relationship holds for the weight enumerator of a nonlinear code. Furthermore, a definition is given for the dual $\mathcal{A}^\perp$ of a nonlinear binary code $\mathcal{A}$ which satisfies $(\mathcal{A}^\perp)^\perp = \mathcal{A}$ provided $\mathcal{A}$ contains the zero codeword.

I. INTRODUCTION

In recent years a number of nonlinear codes have been discovered which have better error-correcting capabilities than any known linear codes (e.g., Refs. 1 and 2). However, very little is known about the properties of such codes. In this paper we study the most basic property, the Hamming weight enumerator (defined in Section II), which gives fundamental information about the error probability when the code is used in various error-correction schemes (Ref. 3, Ch. 16). In 1963 one of us showed that the Hamming and the complete weight enumerators of a linear code are related in a simple way to those of the dual code (Ref. 4; Theorems 1 and 3 below). The requirement that the code be linear is unsatisfactory for two reasons: (*i*) Several pairs of nonlinear

codes $\mathfrak{A}$, $\mathfrak{B}$ are known whose weight enumerators satisfy Theorem 3. One example of such a pair is given by the Preparata[2] and Kerdock[1] codes, another by the code shown in Fig. 1. (*ii*) The important theorem of S. P. Lloyd (giving a necessary condition for the existence of a prefect code) may be deduced for linear codes as a corollary to Theorem 3 (Ref. 4, Lemma 2.15), but may be proved directly without assuming linearity (Ref. 5; Ref. 6, p. 111).

It is the purpose of the present paper, therefore, to define the "weight enumerators of the dual code" so as to make Theorems 1 and 3 (and the corresponding theorem for the Lee weight enumerator, Theorem 2) valid even for nonlinear codes.

Furthermore, if $\mathfrak{A}$ is a nonlinear binary code which contains the zero codeword, we define the formal dual $\mathfrak{A}^\perp$ so as to satisfy:

(*i*) $(\mathfrak{A}^\perp)^\perp = \mathfrak{A}$,
(*ii*) if $\mathfrak{A}$ is linear the two definitions of $\mathfrak{A}^\perp$ agree.

The paper is arranged as follows. Section II states the three MacWilliams identities (Theorems 1, 2, 3). Section III treats the binary case, when the three theorems coincide. The formal dual of a nonlinear binary code is defined in Section 3.5. Section IV treats the general case, first proving Theorem 1 and then deducing Theorems 2, 3 from it. In Section V we discuss properties of the "weights of the dual code" $B(i)$. However, the problem of finding conditions for the $B(i)$ to be positive integers remains unsolved.

0	0	0	0	0	0	0	0		1	1	1	1	1	1	1	1
1	1	0	0	0	0	0	0		0	0	1	1	1	1	1	1
1	0	1	0	0	0	0	0		0	1	0	1	1	1	1	1
1	0	0	1	0	0	0	0		0	1	1	0	1	1	1	1
1	0	0	0	1	0	0	0		0	1	1	1	0	1	1	1
1	0	0	0	0	1	0	0		0	1	1	1	1	0	1	1
1	0	0	0	0	0	1	0		0	1	1	1	1	1	0	1
1	0	0	0	0	0	0	1		0	1	1	1	1	1	1	0

Fig. 1—The sixteen rows form a nonlinear code $\mathfrak{A}$.

II. WEIGHT ENUMERATORS

Let F be a finite field $GF(q)$, where q is a prime power; and let F^n be a vector space of dimension n over F. A *linear code* $\mathfrak{A}$ of length n over $GF(q)$ is a subspace of F^n, and $\mathfrak{A}^\perp$ denotes the orthogonal subspace or *dual* code of $\mathfrak{A}$. A code is *self-dual* if $\mathfrak{A} = \mathfrak{A}^\perp$. A *nonlinear* code is any subset of F^n. In this paper a code is linear unless stated otherwise.

We propose to describe the code vectors of a code $\mathfrak{A}$ in three ways, giving progressively less information (but becoming progressively easier to handle).

2.1 *The Complete Weight Enumerator*

Let the elements of F be $\omega_0 = 0, \omega_1, \omega_2, \cdots, \omega_{q-1}$, in some fixed order. The *composition* of a vector $\mathbf{v} \in F^n$ is defined to be

$$\text{comp } (\mathbf{v}) = \mathbf{s} = (s_0, s_1, \cdots, s_{q-1}), \tag{1}$$

where $s_i = s_i(\mathbf{v})$ is the number of coordinates of $\mathbf{v}$ equal to ω_i. Clearly $\sum_{i=0}^{q-1} s_i = n$.

Let $A(\mathbf{t})$ be the number of vectors $\mathbf{v}$ in $\mathfrak{A}$ with comp $(\mathbf{v}) = \mathbf{t}$. The set of integers $\{A(\mathbf{t})\}$ is the *complete weight enumerator* of $\mathfrak{A}$.

The first MacWilliams identity relates the complete weight enumerators of $\mathfrak{A}$ and $\mathfrak{A}^\perp$. (Ref. 4, Lemma 2.7. See also Refs. 7 and 8.)

Theorem 1: *If $\mathfrak{A}$ is a linear code with complete weight enumerator $\{A(\mathbf{t})\}$, and its dual code $\mathfrak{A}^\perp$ has complete weight enumerator $\{B(\mathbf{t})\}$, then*

$$\sum_{\mathbf{s}} B(\mathbf{s}) z_0^{s_0} \cdots z_{q-1}^{s_{q-1}} = \frac{1}{|\mathfrak{A}|} \sum_{\mathbf{t}} A(\mathbf{t}) \prod_{l=0}^{q-1} \left(\sum_{i=0}^{q-1} \mathfrak{X}(\omega_i \omega_l) z_i \right)^{t_l} \tag{2}$$

where the z_i are indeterminates and $\mathfrak{X}$ is a character on $GF(q)$ (defined in Section 4.2).

2.2 *The Lee Weight Enumerator*

For $q = 2$ this description coincides with the preceding, and for $q = 2^s$, $s > 1$ it is not defined; so in this section q is assumed to be an odd prime power.

For q prime, we wish to classify the coordinates of the code vectors by magnitude. For example, codewords over $GF(5) = \{0, 1, -1, 2, -2\}$ would be classified according to the number of components which are 0, the number which are ± 1, and the number which are ± 2 (but without regard to the actual number which are 1, -1, 2, or -2).

In general, for q a prime power, let the elements of F be $\omega_0 = 0$, $\omega_1, \cdots, \omega_\delta, \omega_{-\delta}, \omega_{-\delta+1}, \cdots, \omega_{-1}$, where $\omega_{-i} = -\omega_i$ and $\delta = \frac{1}{2}(q-1)$.

Then the *Lee weight* of a vector $\mathbf{v} \in F^n$ is defined to be

$$\text{Lee } (\mathbf{v}) = (l_0, l_1, \cdots, l_\delta),$$

where $l_i = l_i(\mathbf{v})$ is the number of coordinates of $\mathbf{v}$ equal to either ω_i or $-\omega_i$. In the notation of eq. (1),

$$\begin{aligned} l_0(\mathbf{v}) &= s_0(\mathbf{v}) \\ l_i(\mathbf{v}) &= s_i(\mathbf{v}) + s_{-i}(\mathbf{v}) \quad \text{for} \quad i = 1, \cdots, \delta. \end{aligned} \tag{3}$$

Let $A^L(\mathbf{t})$ be the number of vectors $\mathbf{v}$ in $\mathfrak{A}$ with Lee $(\mathbf{v}) = \mathbf{t}$; so that $\{A^L(\mathbf{t})\}$ is the *Lee weight enumerator* of $\mathfrak{A}$.

The second MacWilliams identity relates the Lee weight enumerators of $\mathfrak{A}$ and $\mathfrak{A}^\perp$:

Theorem 2:

$$\sum_{\mathbf{s}} B^L(\mathbf{s}) z_0^{s_0} \cdots z_\delta^{s_\delta}$$

$$= \frac{1}{|\mathfrak{A}|} \sum_{\mathbf{t}} A^L(\mathbf{t}) \prod_{l=0}^{\delta} \left(z_0 + \sum_{i=1}^{\delta} (\mathfrak{X}(\omega_i \omega_l) + \mathfrak{X}(-\omega_i \omega_l)) z_i \right)^{t_l}, \tag{4}$$

where $\{B^L(\mathbf{s})\}$ *is the Lee weight enumerator for* $\mathfrak{A}^\perp$.

(Theorem 2 is believed to be new.) The Lee enumerator is important both because it is an appropriate measure for codes to be used in phase-modulation communication schemes (see Lee, Ref. 9; Berlekamp, Ref. 3, p. 205) and as a compromise in giving much more information than the Hamming enumerator, yet requiring only half as many variables as the complete enumerator.

2.3 *The (Hamming) Weight Enumerator*

For the rest of the paper let q be any prime power.

The (Hamming) *weight* of a vector $\mathbf{v}$, $wt(\mathbf{v})$, is the number of its nonzero coordinates, so that

$$wt(\mathbf{v}) = \sum_{i=1}^{q-1} s_i(\mathbf{v}). \tag{5}$$

Let $\mathfrak{A}$ be a linear code of length n over $GF(q)$, and let $A(i)$ be the number of vectors $\mathbf{v}$ in $\mathfrak{A}$ with $wt(\mathbf{v}) = i$. Then $\{A(i)\}$ is the (Hamming or ordinary) *weight enumerator* of $\mathfrak{A}$. Similarly $\{B(i)\}$ denotes the weight enumerator of the dual code $\mathfrak{A}^\perp$. The third MacWilliams identity (Ref. 4, Theorem 1) relates $\{A(i)\}$ and $\{B(i)\}$:

Theorem 3:

$$\sum_{i=0}^{n} B(i)z^i = \frac{1}{|\mathfrak{A}|} \sum_{i=0}^{n} A(i)(1 + (q-1)z)^{n-i}(1-z)^i. \tag{6}$$

2.4 *An Example*

Let $\mathfrak{A}$ be the self-dual code of length 2 over $GF(5)$ consisting of the code vectors 0 0, 1 2, 2 −1, −2 1, −1 −2.

The complete, Lee, and Hamming weight enumerators are, respectively,

$$A(20000) = A(01100) = A(01001) = A(01010) = A(00011) = 1,$$

$$A^L(200) = 1, \qquad A^L(011) = 4,$$

and

$$A(0) = 1, \qquad A(2) = 4.$$

In this case, $\mathfrak{X}(\omega_i\omega_l) = \alpha^{il}$ where $\alpha = e^{(2\pi i)/5} = \cos 72° + i \sin 72°$. Theorems 1, 2, 3 assert (correctly) that

$$z_0^2 + z_1z_2 + z_2z_{-1} + z_1z_{-2} + z_{-2}z_{-1} = \tfrac{1}{5}[(z_0 + z_1 + z_2 + z_{-2} + z_{-1})^2$$
$$+ (z_0 + \alpha z_1 + \alpha^2 z_2 + \alpha^3 z_{-2} + \alpha^4 z_{-1})(z_0 + \alpha^2 z_1 + \alpha^4 z_2 + \alpha z_{-2} + \alpha^3 z_{-1})$$
$$+ (z_0 + \alpha^2 z_1 + \alpha^4 z_2 + \alpha z_{-2} + \alpha^3 z_{-1})(z_0 + \alpha^4 z_1 + \alpha^3 z_2 + \alpha^2 z_{-2} + \alpha z_{-1})$$
$$+ (z_0 + \alpha z_1 + \alpha^2 z_2 + \alpha^3 z_{-2} + \alpha^4 z_{-1})(z_0 + \alpha^3 z_1 + \alpha z_2 + \alpha^4 z_{-2} + \alpha^2 z_{-1})$$
$$+ (z_0 + \alpha^3 z_1 + \alpha z_2 + \alpha^4 z_{-2} + \alpha^2 z_{-1})(z_0 + \alpha^4 z_1 + \alpha^3 z_2 + \alpha^2 z_{-2} + \alpha z_{-1})],$$

that

$$z_0^2 + 4z_1z_2 = \tfrac{1}{5}[(z_0 + 2z_1 + 2z_2)^2$$
$$+ 4(z_0 + (\alpha + \alpha^4)z_1 + (\alpha^2 + \alpha^3)z_2)(z_0 + (\alpha^2 + \alpha^3)z_1 + (\alpha + \alpha^4)z_2)],$$

and that

$$1 + 4z^2 = \tfrac{1}{5}[(1 + 4z)^2 + 4(1 - z)^2].$$

III. THE BINARY CASE

All the codes in this section are binary, so that Theorems 1 and 2 coincide with Theorem 3.

3.1 *Preliminaries*

Let $F = GF(2)$; let F^n be a vector space of dimension n over F. For purposes of notation we define a group G which is a multiplicative copy of F^n, as follows. Let $x_1, \cdots, x_n$ be indeterminates satisfying $x_i^2 = 1$ and $x_i x_j = x_j x_i$ for $i, j = 1, \cdots, n$. Then G is the multiplicative group consisting of all products $x_1^{v_1} x_2^{v_2} \cdots x_n^{v_n}$ where v_i is 0 or 1. To each vector

$$\mathbf{v} = (v_1, v_2, \cdots, v_n)$$

in F^n we associate the element

$$x^{\mathbf{v}} = x_1^{v_1} x_2^{v_2} \cdots x_n^{v_n}$$

of G. Thus F^n and G are isomorphic, and addition of vectors in F^n corresponds to multiplication in G.

3.2 *Characters*

Let $\mathfrak{X}_{\mathbf{u}}$, $\mathbf{u} \in F^n$, be a character of G given by

$$\mathfrak{X}_{\mathbf{u}}(x^{\mathbf{v}}) = (-1)^a,$$

where $a = \mathbf{u}\mathbf{v}^T$ is the scalar product of $\mathbf{u}$, $\mathbf{v}$ in $GF(2)$.

Let σ_i be the set of vectors of F^n of weight i. Clearly,

$$|\sigma_i| = \binom{n}{i}.$$

Let

$$X_i = \sum_{\mathbf{v} \in \sigma_i} x^{\mathbf{v}}.$$

(For example, $X_1 = x_1 + x_2 + \cdots + x_n$.) X_i is an element of the group algebra QG of G over the field of rational numbers Q.

$\mathfrak{X}_{\mathbf{u}}$ is extended linearly to elements of QG, for example,

$$\mathfrak{X}_{\mathbf{u}}(X_i) = \sum_{\mathbf{v} \in \sigma_i} \mathfrak{X}_{\mathbf{u}}(x^{\mathbf{v}}).$$

Note that $\mathfrak{X}_{\mathbf{u}}(X_i)$ is a rational integer, not an element of $GF(2)$.

Let S_n be the group of all permutations of n symbols, i.e., the group of all $n \times n$ permutation matrices. $\mathbf{v}\pi$ is the vector obtained from $\mathbf{v}$ by multiplying by the permutation matrix π.

Lemma 3.1:

$$\mathfrak{X}_{\mathbf{u}\pi}(x^{\mathbf{v}}) = \mathfrak{X}_{\mathbf{u}}(x^{\mathbf{v}\pi^T}) \quad \textit{for any } \pi \textit{ in } S_n .$$

Proof:

$$\mathfrak{X}_{\mathbf{u}\pi}(x^{\mathbf{v}}) = (-1)^a,$$
$$a = \mathbf{u}\pi\mathbf{v}^T = \mathbf{u}(\mathbf{v}\pi^T)^T. \qquad \text{Q.E.D.}$$

3.3 *Krawtchouk Polynomials*

The *Krawtchouk polynomial* $P_s(i)$ (a polynomial in s) is defined by

$$(1 + z)^{n-s}(1 - z)^s = \sum_{i=0}^{n} P_s(i)z^i, \tag{7}$$

so that

$$P_s(i) = \sum_{r=0}^{\min(i,s)} (-1)^r \binom{s}{r}\binom{n-s}{i-r} \qquad i = 0, \cdots, n. \tag{8}$$

It follows from the definition that

$$\sum_{i=0}^{n} P_s(i) = 2^n \delta_{s,0}. \tag{9}$$

Other properties may be found in Refs. 10 and 11.

Let J_s be the vector with $v_1 = v_2 = \cdots = v_s = 1$ and $v_{s+1} = \cdots = v_n = 0$.

Lemma 3.2: If $\mathbf{u}$ *has weight* s,

$$\mathfrak{X}_{\mathbf{u}}(X_i) = P_s(i).$$

Proof: Since X_i is clearly invariant under any permutation in S_n we may suppose, by (3.1), that $\mathbf{u} = J_s$.

Consider the formal sum

$$\sum_{i=0}^{n} \mathfrak{X}_{J_s}(X_i)z^i = \mathfrak{X}_{J_s}\left(\sum_{i=0}^{n} X_i z^i\right).$$

Now

$$\sum_{i=0}^{n} X_i z^i = \prod_{j=1}^{n} (1 + x_j z), \tag{10}$$

and

$$\mathfrak{X}_{J_s}(1 + x_j z) = \begin{cases} 1 - z & \text{if } j = 1, \cdots, s, \\ 1 + z & \text{if } j = s + 1, \cdots, n. \end{cases}$$

Thus

$$\sum_{i=0}^{n} \mathfrak{X}_{J_s}(X_i)z^i = (1 + z)^{n-s}(1 - z)^s. \qquad \text{Q.E.D.}$$

Lemma 3.3:

$$\binom{n}{i} P_i(s) = \binom{n}{s} P_s(i).$$

Proof: By rearranging the binomial coefficients in eq. (8).

3.4 *Definition of $B(i)$ and Proof of Theorem 3*

Let $\mathcal{A}$ be an arbitrary (linear or nonlinear) code, i.e., any subset of F^n; let $A(i)$ be the number of vectors in $\mathcal{A}$ of weight i. Define

$$\mathbf{A} = \sum_{\mathbf{v} \in \mathcal{A}} x^{\mathbf{v}};$$

$\mathbf{A}$ is an element of QG. Corresponding to $\mathcal{A}$ we define numbers $B(i)$, $i = 0, 1, \cdots, n$, by

$$B(i) = \frac{1}{|\mathcal{A}|} \sum_{u \in \sigma_i} \mathfrak{X}_u(\mathbf{A}). \tag{11}$$

Note that $B(i)$ is a rational number, perhaps negative.

With this definition of $B(i)$ we can now prove the binary version of Theorem 3, as follows. Define

$$\mathbf{A}^\pi = \sum_{\mathbf{v} \in \mathcal{A}} x^{\mathbf{v}\pi}.$$

We average $\mathbf{A}$ over all equivalent codes $\mathbf{A}^\pi$:

Lemma 3.4:

$$\sum_{\pi \in S_n} \mathbf{A}^\pi = \sum_{i=0}^{n} A(i) i! \, (n - i)! \, X_i \, .$$

Proof: Let $\mathbf{v}$ be a vector of weight i in $\mathcal{A}$. The $i!$ permutations of the nonzero symbols of $\mathbf{v}$ leave $\mathbf{v}$ unchanged, as do the $(n - i)!$ permutations of the places in which $\mathbf{v}$ contains zero. Thus

$$\sum_{\pi \in S_n} x^{\mathbf{v}\pi} = i! \, (n - i)! \, X_i \, . \qquad \text{Q.E.D.}$$

Lemma 3.5:

$$B(j) = \frac{1}{|\mathcal{A}|} \frac{1}{j! \, (n - j)!} \sum_{\pi \in S_n} \mathfrak{X}_{J_j \pi}(\mathcal{A}).$$

Proof: As π runs through S_n, $J_j\pi$ runs through $j!(n - j)!$ copies of σ_j.

Q.E.D.

Proof of Theorem 3: By (3.5), (3.1):

$$B(j) = \frac{1}{|\mathfrak{A}|}\frac{1}{j!\,(n-j)!}\mathfrak{X}_{J_i}\Big(\sum_{\pi \in S_n} \mathfrak{A}^{\pi}\Big)$$

$$= \frac{1}{|\mathfrak{A}|}\frac{1}{j!\,(n-j)!}\mathfrak{X}_{J_i}\Big(\sum_{i=0}^{n} A(i)i!\,(n-i)!\,X_i\Big) \quad \text{by (3.4)},$$

$$= \frac{1}{|\mathfrak{A}|}\sum_{i=0}^{n} A(i)\frac{i!\,(n-i)!}{j!\,(n-j)!}P_i(i) \quad \text{by (3.2)},$$

$$= \frac{1}{|\mathfrak{A}|}\sum_{i=0}^{n} A(i)P_i(j) \quad \text{by (3.3)}.$$

Multiply both sides by z^j and sum on j:

$$\sum_{j=0}^{n} B(j)z^j = \frac{1}{|\mathfrak{A}|}\sum_{i=0}^{n} A(i)\sum_{j=0}^{n} P_i(j)z^j$$

$$= \frac{1}{|\mathfrak{A}|}\sum_{i=0}^{n} A(i)(1+z)^{n-i}(1-z)^i. \qquad \text{Q.E.D.}$$

In the next section we show that in the case $\mathfrak{A}$ is linear, $B(i)$ is the usual weight distribution of the dual code.

3.5 *The Dual Code*

If $\mathfrak{A} = \sum_{v \in F^n} \alpha_v x^v$, $\alpha_v \in Q$, is any element of QG for which $A(0) = 1$, we define its formal weight distribution to be $\{A(i)\}$, where

$$A(i) = \sum_{v \in \sigma_i} \alpha_v , \tag{12}$$

$$|\mathfrak{A}| = \sum_{i=0}^{n} A(i), \tag{13}$$

and its formal dual to be

$$\mathfrak{A}^{\perp} = \frac{1}{|\mathfrak{A}|}\sum_{u \in F^n} \mathfrak{X}_u(\mathfrak{A})x^u. \tag{14}$$

It follows from (12) that the formal weight distribution of $\mathfrak{A}^{\perp}$ is $\{B(i)\}$, where

$$B(i) = \frac{1}{|\mathfrak{A}|}\sum_{u \in \sigma_i} \mathfrak{X}_u(\mathfrak{A}). \tag{11'}$$

If $\mathfrak{A}$ is a linear or nonlinear code, then clearly (12), (13) give the usual weight distribution and total number of codewords, and eq. (11') for $B(i)$ coincides with eq. (11) of Section 3.4.

Theorem 4: *If $\mathfrak{C}$ is a linear code, then the expressions (14), (11′) for its dual code and weight distribution of dual code, coincide with the usual definitions.*

Proof: If $\mathbf{u}$ is in the dual subspace to $\mathfrak{C}$, then $\mathfrak{X}_u(x^v) = 1$ for all $\mathbf{v} \in \mathfrak{C}$, so $\mathfrak{X}_u(\mathfrak{C}) = |\mathfrak{C}|$. If $\mathbf{u} \notin \mathfrak{C}^\perp$, then $\mathbf{uv}^T \equiv 1$ (modulo 2) for exactly half the vectors $\mathbf{v} \in \mathfrak{C}$, so

$$\mathfrak{X}_u(\mathfrak{C}) = 0 \quad \text{for} \quad \mathbf{u} \notin \mathfrak{C}^\perp.$$

Therefore from (14),

$$\mathfrak{C}^\perp = \frac{1}{|\mathfrak{C}|} \sum_{u \in \mathfrak{C}^\perp} x^u. \qquad \text{Q.E.D.}$$

Combining Theorem 4 with the results of the last section, we have completed the proof of Theorem 3 for binary linear codes.

Theorem 5: *Let $\mathfrak{C} = \sum_{v \in F^n} \alpha_v x^v$, $\alpha_v \in Q$, be any element of QG for which $A(0) = 1$, with formal dual $\mathfrak{C}^\perp$ given by eq. (14). Then*

(*i*) $|\mathfrak{C}|\,|\mathfrak{C}^\perp| = 2^n$,
(*ii*) $(\mathfrak{C}^\perp)^\perp = \mathfrak{C}$.

(Note that by the earlier remarks this theorem includes linear and nonlinear binary codes as a special case.)

Proof: (*i*) Set $z = 1$ in Theorem 3.
(*ii*) From (14), $(\mathfrak{C}^\perp)^\perp = \sum_{u \in F^n} \beta_u x^u$, where

$$\begin{aligned}
\beta_u &= \frac{1}{|\mathfrak{C}^\perp|} \mathfrak{X}_u(\mathfrak{C}^\perp), \\
&= \frac{1}{2^n} \sum_{v \in F^n} \mathfrak{X}_v(\mathfrak{C}) \mathfrak{X}_u(x^v) \quad \text{by (i), (14)}, \\
&= \frac{1}{2^n} \sum_{v \in F^n} \mathfrak{X}_v\Big(\sum_{w \in F^n} \alpha_w x^w\Big) \mathfrak{X}_u(x^v), \\
&= \frac{1}{2^n} \sum_{w \in F^n} \alpha_w \sum_{v \in F^n} (-1)^{v(u+w)^T}, \\
&= \frac{1}{2^n} (2^n \alpha_u),
\end{aligned}$$

since the innermost sum is zero unless $\mathbf{u} = \mathbf{w}$. Q.E.D.

Remarks: In spite of Theorem 5, eq. (14) is not always a satisfactory definition of the dual of a nonlinear code, even in the binary case.

For example, Fig. 1 shows a nonlinear code with weight distribution $A(0) = A(8) = 1$, $A(2) = A(6) = 7$, and

$$\mathfrak{C} = 1 + x_1(x_2 + x_3 + \cdots + x_8) + x_1 \cdots x_8\left(1 + \frac{1}{x_1}\left(\frac{1}{x_2} + \cdots + \frac{1}{x_8}\right)\right).$$

When the weight distribution is substituted in the right-hand side of the MacWilliams identity (6), $B(i)$ is found to be the same as $A(i)$ (Ref. 4, bottom of p. 82) so that this code is in some sense self-dual. However, although eq. (11) correctly gives the weight distribution $B(0) = B(8) = 1$, $B(2) = B(6) = 7$, eq. (14) gives

$$\mathfrak{C}^{\perp} = 1 - \tfrac{1}{2}x_1(x_2 + x_3 + \cdots + x_8) + \tfrac{1}{2}\sum_{2 \le i < j \le 8} x_i x_j + \cdots$$

which seems unsatisfactory. A better definition of the dual of a nonlinear code has recently been given by P. Delsarte and J. -M. Goethals (private communication).

IV. THE GENERAL CASE

4.1 *Preliminaries*

Let $q = p^f$, $f \geqq 1$, where p is prime; and let $F = GF(q) = \{\omega_0 = 0, \omega_1, \cdots, \omega_{q-1}\}$. Let $x_i^{(\omega_j)}$ be commuting indeterminates satisfying

$$x_i^{(\omega_j)} x_i^{(\omega_k)} = x_i^{(\omega_j + \omega_k)};$$

and let G be the multiplicative group consisting of all products $x_1^{(v_1)} x_2^{(v_2)} \cdots x_n^{(v_n)}$, $v_i \in F$. To each vector $\mathbf{v} = (v_1, \cdots, v_n)$ in F^n we associate the element $x^{(\mathbf{v})} = x_1^{(v_1)} \cdots x_n^{(v_n)}$ of G; as in Section 3.1, G is a multiplicative copy of F^n. Let $\mathfrak{C}G$ be the group algebra of G over the complex numbers.

4.2 *Characters*

Let $p(x)$ be a primitive irreducible polynomial of degree f over $GF(p)$, and let α be a root of $p(x)$. Then any element $\lambda \in GF(q)$ has the canonical representation

$$\lambda = \lambda_0 + \lambda_1\alpha + \lambda_2\alpha^2 + \cdots + \lambda_{f-1}\alpha^{f-1}, \qquad \lambda_i \in GF(p).$$

If $GF(q)$ is considered as an additive group, it forms an abelian group, denoted by $(GF(q), +)$, which is isomorphic to the direct product of f copies of $GF(p)$; the isomorphism being given for example by

$$\lambda \leftrightarrow (\lambda_0, \lambda_1, \cdots, \lambda_{f-1}).$$

A character $\mathfrak{X}$ on $GF(q)$ is a homomorphism from $(GF(q), +)$ to the multiplicative group of the complex numbers. Define a fixed character on $GF(q)$ by

$$\mathfrak{X}(\lambda) = \xi^{\lambda_0}$$

where $\xi = e^{(2\pi i)/p}$, and

$$\mathfrak{X}(\lambda + \mu) = \xi^{\lambda_0 + \mu_0}.$$

All characters on $GF(q)$ are now given by

$$\mathfrak{X}_\nu(\lambda) = \mathfrak{X}(\lambda\nu), \qquad \text{all } \nu \in GF(q).$$

All of the following depends on the choices of $p(x)$, α, and $\mathfrak{X}$; this dependence on coordinatization seems inevitable in studying codes over $GF(q)$.

Define a character $\mathfrak{X}_u$ on G by

$$\mathfrak{X}_u(x^v) = \mathfrak{X}(\mathbf{u}\mathbf{v}^T) = \mathfrak{X}\left(\sum_{i=1}^{n} u_i v_i\right) \tag{15}$$

where $\sum_{i=1}^{n} u_i v_i \in GF(q)$. These characters form a group isomorphic to G (and to F^n): $\mathfrak{X}_u \leftrightarrow x^u$. We extend $\mathfrak{X}_u$ to $\mathfrak{C}G$ by linearity.

Lemma 4.1:

$$\mathfrak{X}_{u\pi}(x^{(v)}) = \mathfrak{X}_u(x^{(v\pi^T)}) \quad \text{for any } \pi \in S_n .$$

The proof is straightforward and is omitted.

4.3 *Generalized Krawtchouk Polynomials.*

Let $\mathbf{s} = (s_0, s_1, \cdots, s_{q-1})$, $\mathbf{t} = (t_0, t_1, \cdots, t_{q-1})$ be compositions as defined in Section 2.1. The *generalized Krawtchouk polynomial* $P_{\mathbf{s}}(\mathbf{t})$ is defined by

$$\prod_{l=0}^{q-1} \left(\sum_{i=0}^{q-1} \mathfrak{X}(\omega_i \omega_l) z_i \right)^{s_l} = \sum_{\mathbf{t}} P_{\mathbf{s}}(\mathbf{t}) z_0^{t_0} z_1^{t_1} \cdots z_{q-1}^{t_{q-1}}. \tag{16}$$

$$\text{Let} \quad X_{\mathbf{t}} = \sum_{\substack{v \in F^n \\ \text{comp } (v) = \mathbf{t}}} x^v.$$

Lemma 4.2:

$$\prod_{k=1}^{n} \sum_{i=0}^{q-1} x_k^{(\omega_i)} z_i = \sum_{\mathbf{t}} X_{\mathbf{t}} z_0^{t_0} z_1^{t_1} \cdots z_{q-1}^{t_{q-1}}.$$

This is a straightforward generalization of eq. (10). For example,

expand the product ($n = 3$, $q = 4$)

$$(x_1^{(\omega_0)}z_0 + x_1^{(\omega_1)}z_1 + x_1^{(\omega_2)}z_2 + x_1^{(\omega_3)}z_3)$$
$$\cdot(x_2^{(\omega_0)}z_0 + x_2^{(\omega_1)}z_1 + x_2^{(\omega_2)}z_2 + x_2^{(\omega_3)}z_3)$$
$$\cdot(x_3^{(\omega_0)}z_0 + x_3^{(\omega_1)}z_1 + x_3^{(\omega_2)}z_2 + x_3^{(\omega_3)}z_3).$$

Lemma 4.3: For any composition $\mathbf{s}$ *let*

$$\overleftarrow{s_0} \longrightarrow \overleftarrow{s_1} \longrightarrow \qquad \overleftarrow{s_{q-1}} \longrightarrow$$

$$\mathbf{u} = (\omega_0\omega_0 \cdots \omega_0\omega_1\omega_1 \cdots \omega_1 \cdots \omega_{q-1}\omega_{q-1} \cdots \omega_{q-1})$$

so that comp $(\mathbf{u}) = \mathbf{s}$. *Then*

$$\mathfrak{X}_{\mathbf{u}}(X_{\mathbf{t}}) = P_{\mathbf{s}}(\mathbf{t}).$$

Proof: Consider the formal sum

$$\sum_{\mathbf{t}} \mathfrak{X}_{\mathbf{u}}(X_{\mathbf{t}})z_0^{t_0} \cdots z_{q-1}^{t_{q-1}} = \mathfrak{X}_{\mathbf{u}}\left(\prod_{k=0}^{n}\sum_{i=0}^{q-1} x_k^{(\omega_i)}z_i\right) \text{ by (4.2)},$$

$$= \prod_{k=1}^{n}\sum_{i=0}^{q-1} \mathfrak{X}_{\mathbf{u}}(x_k^{(\omega_i)})z_i$$

$$= \prod_{k=1}^{n}\sum_{i=0}^{q-1} \mathfrak{X}(u_k\omega_i)z_i \text{ by eq. (15)},$$

$$= \prod_{l=0}^{q-1}\left(\sum_{i=0}^{q-1} \mathfrak{X}(\omega_i\omega_l)z_i\right)^{s_l} \text{ by the form of } \mathbf{u},$$

$$= \sum_{\mathbf{t}} P_{\mathbf{s}}(\mathbf{t})z_0^{t_0} \cdots z_{q-1}^{t_{q-1}}$$

by eq. (16). Q.E.D.

For a composition $\mathbf{s}$, let $\binom{n}{\mathbf{s}}$ denote the multinomial coefficient $n!/(s_0!s_1! \cdots s_{q-1}!)$.

Lemma 4.4:

$$\binom{n}{\mathbf{s}}P_{\mathbf{s}}(\mathbf{t}) = \binom{n}{\mathbf{t}}P_{\mathbf{t}}(\mathbf{s}).$$

Proof: Set $\alpha_l = \sum_{i=0}^{q-1} \mathfrak{X}(\omega_i\omega_t)z_i$, so (16) becomes

$$\prod_{l=0}^{q-1} \alpha_l^{s_l} = \sum_{\mathbf{t}} P_{\mathbf{s}}(\mathbf{t}) \prod_i z_i^{t_i}.$$

Multiply by $\prod_{l=0}^{q-1} \binom{n}{\mathbf{s}} y_l^{s_l}$ and sum on $\mathbf{s}$:

$$\sum_{\mathbf{s}} \binom{n}{\mathbf{s}} \prod_{l=0}^{q-1} (\alpha_l y_l)^{s_l} = \sum_{\mathbf{s},\mathbf{t}} \binom{n}{\mathbf{s}} P_{\mathbf{s}}(\mathbf{t}) \prod_i z_i^{t_i} \prod_l y_l^{s_l}. \qquad (17)$$

The left-hand side is

$$(\alpha_0 y_0 + \alpha_1 y_1 + \cdots + \alpha_{q-1} y_{q-1})^n$$

which rearranged becomes

$$(\beta_0 z_0 + \beta_1 z_1 + \cdots + \beta_{q-1} y_{q-1})^n, \tag{18}$$

where

$$\beta_i = \sum_{l=0}^{q-1} \mathfrak{X}(\omega_i \omega_l) y_l \ .$$

Expanding (18) we get

$$\sum_{\mathbf{t}} \binom{n}{\mathbf{t}} \prod_{i=0}^{q-1} \left(\sum_{l=0}^{q-1} \mathfrak{X}(\omega_i \omega_l) y_l \right)^{t_i} z_i^{t_i} = \sum_{\mathbf{s},\mathbf{t}} \binom{n}{\mathbf{t}} \sum_{\mathbf{s}} P_{\mathbf{t}}(\mathbf{s}) \prod_l y_l^{s_l} \prod_i z_i^{t_i}. \tag{19}$$

Equating coefficients in (17), (19) gives the result. Q.E.D.

4.4 *Definition of $B(\mathbf{s})$ and Proof of Theorem 1*

As in Section 2.1, let $\mathfrak{A}$ be any code in F^n, with complete weight enumerator $\{A(\mathbf{t})\}$; and let

$$\mathfrak{A} = \sum_{\mathbf{v} \in \mathfrak{A}} x^{\mathbf{v}}$$

be the corresponding element of $\mathfrak{C}G$. For each composition $\mathbf{s}$ define

$$B(\mathbf{s}) = \frac{1}{|\mathfrak{A}|} \sum_{\substack{\mathbf{u} \in F^n \\ \text{comp } (\mathbf{u}) = \mathbf{s}}} \mathfrak{X}_{\mathbf{u}}(\mathfrak{A}). \tag{20}$$

In general $B(\mathbf{s})$ is a complex number. With this definition of $B(\mathbf{s})$ we can now prove Theorem 1.

Remark: If $\mathfrak{A}$ is a linear code it follows immediately (as in the proof of Theorem 4) that $\{B(\mathbf{s})\}$ is the composition of the dual code to $\mathfrak{A}$.

We first average $\mathfrak{A}$ over all equivalent codes. For a vector $\mathbf{u}$ of composition t,

$$\sum_{\pi \in S_n} x^{\mathbf{u}\pi} = \prod_{i=0}^{q-1} (t_i!) \sum_{\substack{\mathbf{v} \in F^n \\ \text{comp } (\mathbf{v}) = \mathbf{t}}} x^{\mathbf{v}}.$$

Set $d(\mathbf{t}) = \prod_{i=1}^{q-1} (t_i!)$. Then

$$\sum_{\pi \in S_n} \mathfrak{A}^{\pi} = \sum_{\mathbf{t}} d(\mathbf{t}) A(\mathbf{t}) X_{\mathbf{t}} \ . \tag{21}$$

Proof of Theorem 1:

From eq. (20),

$$|\mathcal{C}|\, B(\mathbf{s}) = \sum_{\substack{\mathbf{u} \in F^n \\ \text{comp}(\mathbf{u}) = \mathbf{s}}} \mathfrak{X}_{\mathbf{u}}(\mathcal{C})$$

$$= \frac{1}{d(\mathbf{s})} \sum_{\pi \in S_n} \mathfrak{X}_{\mathbf{u}\pi}(\mathcal{C})$$

$$= \frac{1}{d(\mathbf{s})} \mathfrak{X}_{\mathbf{u}}\Big(\sum_{\pi \in S_n} \mathcal{C}^{\pi}\Big) \quad \text{by (4.1)},$$

[$\mathbf{u}$ is now the vector defined in Lemma (4.3)],

$$= \frac{1}{d(\mathbf{s})} \sum_{\mathbf{t}} d(\mathbf{t}) A(\mathbf{t}) \mathfrak{X}_{\mathbf{u}}(X_{\mathbf{t}}) \qquad \text{by (21)},$$

$$= \sum_{\mathbf{t}} \frac{d(\mathbf{t})}{d(\mathbf{s})} A(\mathbf{t}) P_{\mathbf{s}}(\mathbf{t}) \qquad \text{by (4.3)},$$

$$= \sum_{\mathbf{t}} P_{\mathbf{t}}(\mathbf{s}) A(\mathbf{t}) \qquad \text{by (4.4)}.$$

Multiply both sides by $z_0^{s_0} \cdots z_{q-1}^{s_{q-1}}$ and sum over all compositions $\mathbf{s}$.

Q.E.D.

4.5 *Proofs of Theorems 2 and 3.*

We use the notation of Sections 2.2 and 2.3.

Proof of Theorem 2:

In eq. (2) replace z_i by z_i for $1 \leqq i \leqq \delta$. Then using eq. (3), we see that eq. (2) collapses into eq. (4). Q.E.D.

Proof of Theorem 3:

In eq. (2) set $z_0 = 1$, $z_i = z$ for $i \neq 0$, and use eq. (5) to obtain (6).

Q.E.D.

V. DISCUSSION

We return to the binary case, which is easier to visualize.

The Hamming distance between vectors $\mathbf{u}$, $\mathbf{v}$ is the weight of $\mathbf{u} + \mathbf{v}$ (the weight of $\mathbf{u} - \mathbf{v}$ if not binary). Coding theorists are interested in the distance structure of a code, not just in its weight structure. For linear codes, these are the same; they may also be the same for nonlinear codes, as in the example in Fig. 1. The following lemma is obvious.

Lemma 5.1: The distance and weight structure of a code $\mathcal{A}$ are the same if and only if the weight structure of $\mathcal{A} + \mathbf{v}$ is the same as that of $\mathcal{A}$ for all $\mathbf{v} \in \mathcal{A}$.

A code of this type will be said to have property 5.1. From now on we restrict ourselves to such codes.

A code with property 5.1 clearly contains the vector $\mathbf{0}$. The element of QG corresponding to $\mathcal{A} + \mathbf{v}$ is $\mathcal{A}x^{\mathbf{v}}$.

Property 5.1 implies that

$$|\mathcal{A}|\, B(s) = \sum_{\mathbf{u} \in \sigma_s} \mathfrak{X}_{\mathbf{u}}(\mathcal{A}) = \sum_{\mathbf{u} \in \sigma_s} \mathfrak{X}_{\mathbf{u}}(\mathcal{A}x^{\mathbf{v}}) \quad \text{for} \quad \mathbf{v} \in \mathcal{A}.$$

Lemma 5.2: Property 5.1 implies that $B(s) \geqq 0$.

Proof: Take the sum over all $\mathbf{v} \in \mathcal{A}$ of the equation

$$\begin{aligned} |\mathcal{A}|\, B(s) &= \sum_{\mathbf{u} \in \sigma_s} \mathfrak{X}_{\mathbf{u}}(\mathcal{A}x^{\mathbf{v}}). \\ |\mathcal{A}|^2 B(s) &= \sum_{\mathbf{u} \in \sigma_s} \mathfrak{X}_{\mathbf{u}} \sum_{\mathbf{v} \in \mathcal{A}} (\mathcal{A}x^{\mathbf{v}}) \\ &= \sum_{\mathbf{u} \in \sigma_s} \mathfrak{X}_{\mathbf{u}}(\mathcal{A}) \sum_{\mathbf{v} \in \mathcal{A}} \mathfrak{X}_{\mathbf{u}}(x^{\mathbf{v}}) \\ &= \sum_{\mathbf{u} \in \sigma_s} (\mathfrak{X}_{\mathbf{u}}(\mathcal{A}))^2. \end{aligned}$$

Q.E.D.

Corollary 5.3: If $B(s) = 0$ then $\mathfrak{X}_{\mathbf{u}}(\mathcal{A}) = 0$ for each $\mathbf{u} \in \sigma_s$.

Property 5.1 does not imply that $B(s)$ is an integer. Since by Theorem 5, $\sum_s B(s) = 2^n/|\mathcal{A}|$, $B(s)$ cannot all be integers unless $|\mathcal{A}| = 2^k$. For example, the code $\begin{pmatrix} 000 \\ 110 \\ 011 \end{pmatrix}$ has property 5.1, but the $B(s)$ are not all integers.

At present we have a satisfactory interpretation for $A(s)$, $B(s)$ if $\sum_{\pi \in S_n} \mathcal{A}^{\pi}$ can be generated by a linear code. ($\mathcal{A}$ need not be linear; any collection of vectors with the same weights as the vectors of a linear code will give the same average.) It would be very desirable to find an explanation for the cases in which $A(s)$, $B(s)$ can be thought of as the weight distribution of nonlinear codes.

VI. ACKNOWLEDGMENT

The authors would like to state that they found the basic idea of this paper, and much more, in J. H. van Lint's book *Coding Theory*.[6]

Added to galley proof:

Since this paper was written, it has come to our attention that Neal Zierler (unpublished) discovered the nonlinear MacWilliams identity for Hamming weight enumerators in 1966.

REFERENCES

1. Kerdock, A. M., "A Class of Low-Rate Nonlinear Codes," to appear in Info. and Control.
2. Preparata, F. P., "A Class of Optimum Nonlinear Double-Error Correcting Codes," Info. and Control, *13*, No. 4 (October 1968), pp. 378–400.
3. Berlekamp, E. R., *Algebraic Coding Theory*, New York: McGraw-Hill, 1968.
4. MacWilliams, F. J., "A Theorem on the Distribution of Weights in a Systematic Code," B.S.T.J., *42*, No. 1 (Janaury 1963), pp. 79–94.
5. Lloyd, S. P., "Binary Block Coding," B.S.T.J., *36*, No. 2 (March 1956), pp. 517–535.
6. van Lint, J. H., *Coding Theory*, New York: Springer-Verlag, 1971.
7. Assmus, E. F., Jr., "Research to Develop the Algrbraic Theory of Codes," Sylvania Electronic Systems, Waltham, Mass., Report AFCRL-67-0365, June 1967; especially Part V.
8. Gleason, A. M., "Weight Polynomials of Self-Dual Codes and the MacWilliams Identities," Actes, Congrès intern. Math., 1970, Vol. 3, pp. 211–215; Paris: Gauthier-Villars, 1971.
9. Lee, C. Y., "Some Properties of Non-binary Error-Correcting Codes," IEEE Trans. Info. Theory, *IT-4*, No. 2 (June 1958), pp. 77–82.
10. Krawtchouk, M., "Sur une généralisation des polynomes d'Hermite," Comptes Rendus, *189*, 1929, pp. 620–622.
11. Szegö, G., *Orthogonal Polynomials*, Colloquium Publications, Vol. 23, New York: American Mathematical Society, revised edition, 1959, pp. 35–37.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Stochastic Stability of Delta Modulation

By ALLEN GERSHO

(Manuscript received December 3, 1971)

The discrete-time model of delta modulation is considered for a stationary random input process with a rational spectral density, and an autocovariance that goes to zero as the lag approaches infinity. For leaky integration, the joint distribution of input and decoded approximation processes is shown to approach a unique stationary distribution from any initial condition. Under the stationary distribution, the decoded process may take on all values in a bounded interval that is independent of the input process. For the often-studied ideal integration model of delta modulation, it is shown that the successive distributions at even parity time instants converge to a limiting stationary distribution, while at odd parity time instants the distributions converge to a different limiting distribution. Under these limiting distributions, the decoded process is assigned a positive probability for each level of a (discrete) lattice of amplitudes. The mean-absolute approximation error and mean-absolute amplitude of the decoded process are shown to be finite under the limiting distributions. For both ideal and leaky integration cases, an explicit upper bound on mean-absolute approximation error is given, which is independent of the spectral density of the input process.

I. INTRODUCTION

In spite of the great simplicity of delta modulation as an analog-to-digital encoding technique, it has not yet succumbed to an adequate mathematical analysis. Although realistic inputs such as speech are extremely difficult to characterize, considerable insight could be obtained from a thorough analysis for the case of a stationary random input process with a prescribed spectral density. Yet no such results have been obtained because of the mathematical complexity of the nonlinear feedback loop. In fact, the presence of a feedback loop raises the possibility that instability in some sense could arise. The possibility that the decoded signal could "run away" or become unbounded, failing to track the original signal, has never been theoretically excluded.

Although experience with delta modulation shows that such an extreme form of instability never arises, it has never been shown analytically that the mean-square or mean-absolute quantizing noise has a finite upper bound. Another possibility which has never been theoretically excluded is erratic operation, where the statistical average of the quantizing noise magnitude continues to vary with time. In other words, although the input process is stationary, the decoded approximation process would be nonstationary with a time-varying probability distribution even after low-pass filtering. If this were the case, the decoded process would not be replicating the original process very effectively.

Recently, D. Slepian[1] has developed an exact computational approach for finding the joint probability distribution of the original and encoded processes. These results make it possible to accurately compute such curves as the mean-square quantizing noise versus step size for particular spectral densities of the input process. Slepian's results are based on the initial assumption that, for a stationary input process with rational spectral density, the joint probability distribution will approach a unique stationary distribution from any starting condition. (For delta modulation with ideal integration, the stationary distribution actually refers to half the sum of the distributions at two successive time instants to account for the well-known parity change between even and odd amplitude levels.) On a practical level, this stationarity assumption seems to be entirely reasonable; yet it has never been theoretically justified.

Other authors have also assumed stationarity. In particular, for ideal integration H. van de Weg[2] assumed implicitly that the decoded process had two different stationary distributions for the even and odd parity time instants. D. J. Goodman[3] assumed a random phase initial condition so that only one stationary distribution, half the sum of the even and odd parity distributions, need be considered. In both cases the assumption is implicitly made that for any initial condition the delta modulation process will approach a steady-state mode of operation with a separate stationary distribution for even and odd time instants.

As a final argument to point out the need for an analysis of stochastic stability properties of delta modulation, consider the fact that most heuristic and semianalytical approximate considerations of delta modulation are based on the model of an ideal integrator in the feedback path, while most physical realizations involve a leaky integrator. There is a basic qualitative difference between these two cases, even for extremely wide-band integrators. This is because ideal integration gives

equal weight to a current input sample and an arbitrarily remote past input sample, while leaky integration forgets remote past samples. Hence, it is not clear that the ideal integration model is meaningful, even if it is known that the leaky integration system is well behaved. To justify the validity of using an ideal integration model, a rigorous demonstration of the stability of this model is needed.

This paper demonstrates the stochastic stability of delta modulation for both ideal and leaky integration by giving a mathematical proof that the joint distribution of the input and decoded processes approaches a unique stationary distribution from an arbitrary starting point and by deriving an explicit, finite upper bound on the mean-absolute quantizing noise. The input process is assumed to be stationary and continuously distributed, with finite variance, a rational spectral density, and an autocovariance that approaches zero as the lag goes to infinity. A shaping filter with white noise input is assumed as the generating mechanism of the input process.

II. PROBLEM FORMULATION AND SUMMARY OF RESULTS

For most purposes the following discrete-time model of delta modulation is an acceptable description of the actual continuous-time operation.[4] Let u_t denote the sampled analog values of the *input* process at successive time instants $t = 0, 1, 2, \cdots$. The time scale is normalized for convenience without loss of generality. The delta modulator shown in Fig. 1 generates a binary-valued process b_k according to

$$b_k = \operatorname{sgn}(u_k - x_k) \tag{1}$$

where $\operatorname{sgn} y = +1$ if $y \geqq 0$ and -1 if $y < 0$, and x_k is the *decoded process* which approximates u_k. The decoded process x_k is given recursively by

$$x_{k+1} = x_k + \Delta \operatorname{sgn}(u_k - x_k) \tag{2}$$

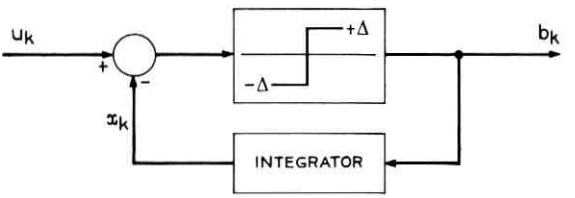

Fig. 1—Delta modulator.

for the *ideal integration* case, and by

$$x_{k+1} = \alpha x_k + \Delta \operatorname{sgn} (u_k - x_k) \tag{3}$$

for the *leaky integration* case with a simple RC integrator, where $\Delta > 0$ is the *step size* and $\frac{1}{2} \leqq \alpha < 1$. (In practice, α is very close to one because the integrator time constant is much longer than the sampling period.) The *quantizing noise*,

$$e_k = u_k - x_k ,$$

is the error at time k due to the analog-to-digital-to-analog processing of the delta modulation system.

Assume the input process u_k is stationary, continuously distributed with finite variance, and has autocovariance approaching zero as the time lag goes to infinity. For convenience, assume also that the probability density of u_k is everywhere positive so that, as in the Gaussian case, there is a positive probability of u_k lying in any open interval. Note that u_k is not assumed to have zero mean value.

A summary of the results obtained in this paper follows. Let $\hat{F}_k$ denote the joint distribution of the input and decoded processes u_k and x_k at time k.

2.1 *Ideal Integration*

(*i*) For any initial condition of the form $x_1 = m\Delta + \theta$ with m an integer and $| \theta | \leqq \Delta/2$, the two sequences of distributions $\{\hat{F}_{2k}\}$ and $\{\hat{F}_{2k+1}\}$ separately converge to unique stationary distributions $\hat{G}_0$ and $\hat{G}_1$, respectively. One distribution assigns positive probability for the process $x_k - \theta$ to even integer multiples of Δ, the other distribution to odd integer multiples, depending on the even-odd parity of the initial integer m.

(*ii*) With these stationary distributions the mean-absolute quantizing noise, averaged over two successive time instants, is bounded according to

$$\tfrac{1}{2}[E \mid x_k - u_k \mid + E \mid x_{k+1} - u_{k+1} \mid] \leqq E \mid u_i \mid + \Delta/2 + 2 \mid \theta \mid . \tag{4}$$

2.2 *Leaky Integration*

(*i*) For any initial value of x_1, the distributions $\{\hat{F}_k\}$ converge to the unique stationary distribution $\hat{G}$, under which x_k may take on all values in the range

$$| x_k | < \frac{\Delta}{1 - \alpha} . \tag{5}$$

(*ii*) Under this stationary distribution, the mean-absolute quantizing noise is bounded according to:

$$E \mid x_k - u_k \mid \leqq E \mid u_i \mid + \Delta/2\alpha. \tag{6}$$

Note the qualitative difference between the leaky and ideal integration cases. In the leaky case, x_k is distributed over a finite interval; in the ideal case, x_k is discretely distributed on a lattice.

III. MARKOVIAN MODELING

Since the input process u_k is stationary with finite variance and rational spectrum, then $\tilde{u}_k = u_k - \mu$ (where μ is the mean value of u_k) can be modeled[5] as the response of a stable discrete-time shaping filter to a zero mean, finite variance "white noise" process w_k (with w_k independent of w_{k-i} for $i = 1, 2, 3, \cdots$). More precisely, a white process w_k and a stable rational shaping filter $H(z)$ can always be specified in such a way that the response $\hat{u}_k$ of the filter to the excitation w_k will be a stationary random process with spectral density identical to that of $\tilde{u}_k$. If w_k is also chosen to be continuously distributed with a positive density, then $\hat{u}_k$ will also satisfy this property. Thus, all the assumptions made in Section II about the process u_k are possessed by the process $\hat{u}_k + \mu$. It is therefore reasonable to study the effect of the delta modulation system for the input $\hat{u}_k + \mu$, whose structure or generating mechanism is known. For the remainder of this paper, no distinction will be made between $\tilde{u}_k$ and $\hat{u}_k$.

Using this model and the assumption that the autocovariance of u_k goes to zero as the lag approaches infinity, Appendix A shows that u_k can be imbedded in a vector Markov process, $\mathbf{d}_k$, with

$$\mathbf{d}_k = (d_{k1}, d_{k2}, \cdots, d_{kn})'$$

and

$$u_k = d_{k1} + \mu \tag{7}$$

where μ denotes the dc value of the input process and n is the number of poles in the shaping filter. The vector $\mathbf{d}_k$ characterizes the state of the filter at time k and is generated by the recursion

$$\mathbf{d}_{k+1} = A\mathbf{d}_k + \mathbf{b}w_k \tag{8}$$

where A is an $n \times n$ matrix with eigenvalues within the unit circle, and $\mathbf{b}$ is a fixed vector. The process $\mathbf{d}_k$ is Markovian, since the conditional distribution of $\mathbf{d}_{k+1}$ given all past states $\mathbf{d}_k, \mathbf{d}_{k-1}, \cdots$, depends

only on the given value of $\mathbf{d}_k$. Appendix B shows that for any initial state $\mathbf{d}_0$, the distribution of $\mathbf{d}_k$ approaches a unique stationary distribution. Equations (2), (7), and (8) for the ideal integrator case, or eqs. (3), (7), and (8) for the leaky integrator case, jointly characterize the evolution in time of a Markovian process whose state $\mathbf{s}_k$ at time k is given by the $n + 1$ component vector

$$\mathbf{s}_k = (x_k , u_k , d_{k2} , d_{k3} , \cdots , d_{kn})'.$$

Then, given the value of $\mathbf{s}_k$, the distribution of $\mathbf{s}_{k+1}$ is completely determined. Henceforth, a *distribution* F_k , describing the joint distribution of the $n + 1$ components of the vector $\mathbf{s}_k$, will be regarded as a set function which assigns a probability $F_k(A)$ to the region A of the $n + 1$ dimensional space of possible values of the state vector $\mathbf{s}_k$. The *probability transition function*[6] characterizing the Markov process is defined as

$$p(\mathbf{s}, A) = P\{\mathbf{s}_{k+1} \varepsilon A \mid \mathbf{s}_k = \mathbf{s}\}$$

which is independent of k. By averaging this conditional probability over a distribution F_k assigned to $\mathbf{s}_k$, the unconditioned distribution F_{k+1} of $\mathbf{s}_{k+1}$ is obtained:

$$F_{k+1}(A) = \int p(\mathbf{s}, A)F_k(d\mathbf{s}) = E_{F_k}p(\mathbf{s}, A). \tag{9}$$

Thus, F_{k+1} is related to F_k by a linear mapping T, so that in operator notation

$$F_{k+1} = TF_k . \tag{10}$$

Note that T plays the same role as the probability transition matrix in Markov chains. The process has a *stationary distribution* G, if $G = TG$, so that G is self-reproducing. If any state vector has distribution G, all subsequent state vectors will have this distribution. The existence of a stationary distribution is a necessary but not sufficient condition for the convergence of the distributions F_{k+1} to a limiting distribution.

The Markovian model will be used in Sections V and VI to obtain the convergence properties of the distributions $\{F_k\}$. But first, it is necessary to obtain a bound on the time and ensemble average of the quantizing noise.

IV. BOUNDING THE TIME-AVERAGED MEAN-ABSOLUTE QUANTIZING NOISE

Suppose that at the initial time instant $k = 1$, the system has an arbitrary initial state $\mathbf{s}_1$. Appendix B shows that for any initial state

$\mathbf{d}_1$ of the shaping filter, the process $\mathbf{d}_k$, and hence u_k, converge in distribution and $E \mid u_k \mid$ converges to c, the mean-absolute value under the stationary distribution of the process u_t.

Squaring both sides of eq. (3) gives

$$\begin{aligned} x_{k+1}^2 &= \alpha^2 x_k^2 - 2\alpha\Delta \mid u_k - x_k \mid + 2\alpha\Delta u_k \operatorname{sgn}(u_k - x_k) + \Delta^2 \\ &\leqq x_k^2 - 2\alpha\Delta \mid u_k - x_k \mid + 2\alpha\Delta \mid u_k \mid + \Delta^2. \end{aligned}$$

Taking the expected value of both sides yields

$$Ex_{k+1}^2 \leqq Ex_k^2 - 2\alpha\Delta E \mid e_k \mid + 2\alpha\Delta E \mid u_k \mid + \Delta^2$$

and iterating backwards gives

$$Ex_{k+1}^2 \leqq x_1^2 + 2\alpha\Delta \sum_{i=1}^{k} (E \mid u_i \mid - E \mid e_i \mid) + \Delta^2 k.$$

Since the left side is nonnegative, it follows that

$$\frac{1}{k} \sum_{i=1}^{k} E \mid e_i \mid \leqq x_1^2/2\alpha\Delta k + \frac{1}{k} \sum_{i=1}^{k} E \mid u_i \mid + \Delta/2\alpha. \tag{11}$$

Hence, using the fact that $E \mid u_i \mid \to c$,

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} E \mid e_i \mid \leqq c + \Delta/2\alpha, \tag{12}$$

which shows that the long-term time average of the mean-absolute quantizing error is bounded. Since the preceding derivation holds for $\alpha \leqq 1$, it applies for both leaky and ideal integration, setting $\alpha = 1$ for ideal integration.

For leaky integration, the decoded process is, in fact, bounded deterministically. Integrating eq. (3) backwards yields

$$x_{k+1} = \alpha^k x_1 + \Delta \sum_{i=1}^{k} \alpha^{k-i} b_i \tag{13}$$

so that

$$\mid x_{k+1} \mid \leqq \alpha^k \mid x_1 \mid + \frac{\Delta}{1-\alpha}(1 - \alpha^k). \tag{14}$$

Hence,

$$\limsup_{k\to\infty} \mid x_k \mid \leqq \frac{\Delta}{1-\alpha}. \tag{15}$$

Thus, the decoded process is bounded with probability one in the case of leaky integration.

V. EXISTENCE OF A STATIONARY DISTRIBUTION

A sequence of random vectors and the corresponding sequence of distributions are said to be *stochastically bounded* if, for any probability ϵ, however small, there is a sufficiently large distance R such that each random vector of the sequence has probability less than ϵ of having length greater than R. Hence, the successive vectors cannot have a positive probability of moving out toward infinity.

For a sequence of distributions $\{F_k\}$, define the associated sequence of *averaged distribution* $\{G_k\}$ by

$$G_k(A) = \frac{1}{k} \sum_{i=1}^{k} F_i(A). \tag{16}$$

Thus, if I is a randomly selected time instant from the first k integers each having equal probability, then $G_k(A)$ is the probability that the random vector $\mathbf{s}_I$ lies in a region A, where $\mathbf{s}_i$ has distribution F_i. If the sequence $\{F_i\}$ is stochastically bounded, then the averaged distributions $\{G_k\}$ are also stochastically bounded; however, the converse is not always true.

To show that the Markov process $\mathbf{s}_k$ defined in Section III has a stationary distribution G, the following theorem, proved in Appendix C, may be used.

Theorem: A Markov process has a stationary distribution if

(*i*) for any initial state $\mathbf{s}_1$, the averaged distributions G_k are stochastically bounded, and

(*ii*) for any region A, let D be the set of points $\mathbf{s}$ at which the transition probability function $p(\mathbf{s}, A)$ is discontinuous and let N_δ be the set of all points whose distance from D is less than δ; then there is a function $C(\delta)$ independent of $\mathbf{s}$ with $C(\delta) \to 0$ as $\delta \to 0$ and for all $\mathbf{s}$,

$$p(\mathbf{s}, N_\delta) \leqq C(\delta). \tag{17}$$

Condition (*i*) excludes the possibility that successive state vectors can move out toward infinity. Condition (*ii*) is concerned with the region of state space where an arbitrarily small perturbation of a given state vector can cause a substantial change in the induced distribution of the state vector at the next time instant. This region is contained within the set N_δ for each $\delta > 0$. The condition requires that this region be suitably unimportant.

To show that the delta modulation process satisfies condition (*i*),

observe that

$$E \mid x_i \mid \leqq E \mid u_i - x_i \mid + E \mid u_i \mid$$

and since $E \mid u_i \mid$ is bounded, eq. (11) shows that for some constant K,

$$\frac{1}{k} \sum_{i=1}^{k} E \mid x_i \mid < \mathrm{K}. \tag{18}$$

Now Chebyshev's inequality,

$$MP \{\mid x_i \mid > M\} \leqq E \mid x_i \mid ,$$

applied to eq. (18) yields

$$\frac{1}{k} \sum_{i=1}^{k} P\{\mid x_i \mid > M\} \leqq \mathrm{K}/M$$

for every $M > 0$, which shows that the averaged distributions of the decoded process x_k are stochastically bounded. But from Appendix B, the $\mathbf{d}_k$ process is stochastically bounded. Hence, the averaged distributions of $\mathbf{d}_k$ are also stochastically bounded. Thus, the marginal distributions of the joint distributions G_n are stochastically bounded so that G_n is itself a stochastically bounded sequence, and condition (i) holds for the ideal integration case. For the leaky integration case, condition (i) is satisfied since eq. (15) shows that the x components of the vectors $\mathbf{s}_k$ are uniformly bounded with probability one, so that the above argument shows that the joint distributions G_n are stochastically bounded.

To verify condition (ii), note from eqs. (2) or (3) that for any region A, $p(\mathbf{s}, A)$ is continuous, except in the set D of all points $\mathbf{s}$ the first two components of which, x and u, are equal. Appendix D shows that given $\mathbf{s}_k$, the variate u_{k+1} is continuously distributed and that this implies that there exists a function $C(\delta)$ which goes to zero as δ approaches zero and

$$P\{\mid x - u_k \mid \mid \mathbf{s}_k\} \leqq C(\delta) \tag{19}$$

where $C(\delta)$ is independent of x and $\mathbf{s}_k$. But since x_{k+1} is completely determined by $\mathbf{s}_k$, (19) implies that

$$P\{\mid x_{k+1} - u_{k+1} \mid < \delta \mid \mathbf{s}_k\} \leqq C(\delta). \tag{20}$$

Hence, eq. (17) is satisfied and condition (ii) holds for both ideal and leaky integration. Therefore, a stationary distribution exists.

VI. ALLOWABLE AMPLITUDE VALUES FOR THE DECODED PROCESS

For ideal integration, it is clear from eq. (2) that an initial condition of the form

$$x_1 = m\Delta + \theta$$

with m an integer and $| \theta | < \Delta$, implies that all subsequent amplitude values of the decoded process will be confined to the lattice

$$x_k = l\Delta + \theta \qquad l = 0, \pm 1, \pm 2, \cdots .$$

Since the preceding results did not specify a particular choice of initial condition, it follows that for each θ, a stationary distribution G_θ exists. For convenience, assume $\theta = 0$. No loss of generality will result because, for any θ, the problem can be converted to the $\theta = 0$ case by replacing the input process u_k by $u_k - \theta$ as can be seen from eq. (2). This simply changes the dc value of the input by, at most, one step size Δ.

Under the assumption that u_k has a positive density, it follows from eq. (2) that there is always a positive probability of either increasing or decreasing by Δ in going from x_k to x_{k+1}. This means that *every* integer multiple of Δ must have a positive probability under the stationary distribution, because each level can always be reached from any other level in a finite number of steps. (On the other hand, if u_k were bounded, then eq. (2) shows that the decoded process would get locked into a bounded set of levels from which it would never escape.)

For leaky integration, the situation is strikingly different. The decoded process will get locked into the bounded region

$$X = \{x: -\Delta/(1 - \alpha) \leqq x \leqq \Delta/(1 - \alpha)\}. \tag{21}$$

This may be seen by noting from eq. (3) that if x_k is in X, then x_{k+1} must also be in X. Consequently, once x_i is in X, it will never escape. If x_k is not in X, then eq. (3) shows that

$$| x_{k+1} | < | x_k |$$

and if b_k has the correct polarity,

$$| x_{k+1} | < \alpha | x_k | . \tag{22}$$

Hence the values $| x_{k+i} |$ must decrease monotonically as long as x_{k+i} remains outside of X. Since u_k has a positive density, b_i must have positive probability of having either polarity; which means the stronger inequality eq. (22) must hold at some subsequent time instants. Hence, the process must eventually enter the region X.

Iterating eq. (3) backwards yields

$$x_{k+1} = \alpha^k x_1 + \Delta \sum_{i=0}^{k-1} \alpha^i b_{k-i} . \tag{23}$$

This shows that the initial value x_1 is gradually forgotten and the set

of allowable values for x_k as k goes to infinity, approaches the set

$$W = \{x: x = \Delta(\pm 1 \pm\alpha \pm\alpha^2 \pm \cdots)\} \tag{24}$$

where all possible sequences of polarities are used to generate values of x. Appendix E proves that, for $\alpha \geqq \frac{1}{2}$, the set W *coincides* with the interval X.*

For the leaky integration case, the stationary distribution G clearly must confine the x component of the state vector to the region X. Furthermore, the following argument shows that G assigns a positive probability to every open subinterval $(y - \epsilon, y + \epsilon)$ with y in X and $\epsilon > 0$. Let $\{c_i\}$ be a suitable binary sequence (generated as in Appendix E) satisfying

$$y = \Delta \sum_{i=0}^{\infty} c_i \alpha^i. \tag{25}$$

Pick the integer N large enough so that

$$\alpha^{N+1} x_1 < \epsilon/3, \quad \text{and} \quad \alpha^N \Delta/(1 - \alpha) < \epsilon/3, \tag{26}$$

and consider the event

$$E = \{b_2 = c_{N-1}, b_3 = c_{N-2}, \cdots, b_{N+1} = c_0\},$$

which has a positive probability for any initial state $\mathbf{s}_1$. Then for $k = N + 1$, eq. (23) can be written in the form

$$x_{N+2} - y = \alpha^{N+1} x_1 + \Delta\alpha^N b_1 - \Delta \sum_{i=N}^{\infty} \alpha^i c_i$$

so that, using eq. (26),

$$| x_{N+2} - y | \leqq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,$$

for x_1 in X. Hence,

$$P\{| x_{N+2} - y | < \epsilon \mid \mathbf{s}_1\} \geqq P\{E \mid s_1\} \tag{27}$$

and averaging eq. (27) over the distribution G for the state $\mathbf{s}_1$ shows that

$$P\{| x_{N+2} - y | < \epsilon\} > 0 \tag{28}$$

where x_{N+2} has the marginal distribution of the first component of G.

VII. CONVERGENCE TO THE LIMITING DISTRIBUTIONS

The following specialized and paraphrased version of a theorem due to J. L. Doob[7] is suited to the delta modulation process:

* For $\alpha < \frac{1}{2}$, it can be shown that W does not coincide with X, in fact W is not even dense in X.

Doob's Theorem: Suppose the Markov process has a stationary distribution G and satisfies the conditions:

(*i*) *For any initial state* $\mathbf{s}_1$, *if* A *is a region with* $G(A) = 0$, *then* $F_k(A) \to 0$.

(*ii*) *If a region* A *satisfies* $p(\mathbf{s}, A) = 1$ *for* $\mathbf{s}$ *in* A, *then* $G(A) = 1$.

Then either

$$F_k \to G \quad \textit{for any initial state}$$

in which case the process is aperiodic, or there exist m *disjoint sets* $A_0, A_1, \cdots, A_{m-1}$ *with* $G(\bigcup_0^m A_i) = 1$ *and* $p(\mathbf{s}, A_{i+1}) = 1$ *if* $\mathbf{s}$ *is in* A_i *for* $i = 0, 1, \cdots, m-2$ *and if* $\mathbf{s}$ *is in* A_{m-1}, *then* $p(\mathbf{s}, A_0) = 1$. *In this case, the process is said to be periodic with period* m. *In particular for* $m = 2$, *there exist two distributions* G_0 *and* G_1 *with* $G = \frac{1}{2}(G_0 + G_1)$, $G_1(A_1) = 1$, $G_0(A_0) = 1$, *and for any initial state in* A_1, $F_{2k+1} \to G_1$ *and* $F_{2k} \to G_0$, *while for any initial state in* A_0, $F_{2k+1} \to G_0$ *and* $F_{2k} \to G_1$.

For ideal integration, Section VI shows that there are no transient levels for the x_k process and Appendix A shows that d_k has a positive probability of lying in any region of n space with nonzero volume. Hence, there is no transient set A with $G(A) = 0$ except for trivial sets with $F_k(A) = 0$ for each k. For leaky integration the only nontrivial sets A are regions where the x component lies outside of X and for such regions, Section VI shows that $F_k(A) \to 0$. Therefore, condition (*i*) is satisfied for both types of integration.

Furthermore, the ergodicity requirement (*ii*) is also seen to be satisfied for both ideal and leaky integration from the results of Section VI.

For ideal integration the process clearly has period 2, since, if A_0 is the set of all state vectors with x components taking on even integer multiples of Δ, and A_1 is the complementary set, then $A_0 \cup A_1$ has probability one under the limiting distribution and the transition probability function has the requisite property implying the state vector alternates between A_0 and A_1. For leaky integration the process is aperiodic since eq. (27) holds for all N sufficiently large so the process cannot satisfy the requirements for periodicity, hence $F_k \to G$. Since a sequence of distributions cannot at the same time converge to two different distributions, it follows that the stationary distribution G is unique for both ideal and leaky integration.

VIII. BOUNDING STEADY-STATE QUANTIZING NOISE

The fact that a sequence of distributions converges to a limiting distribution does not imply that moment functions such as mean or

variance converge to the corresponding moment of the limiting distribution. However, it does imply that the sequence of expectations of a bounded continuous function converge to the expectation of that function under the limiting distribution. It turns out that this property is sufficient to obtain a bound on the least mean-absolute quantizing error under the stationary distributions.

For ideal integration, eq. (12) can be rewritten in the form

$$\limsup_{m\to\infty} \frac{1}{2m} \sum_{i=1}^{m} E(|\, e_{2k} \,| + |\, e_{2k+1} \,|) \leqq \mathrm{K} \tag{29}$$

where $\mathrm{K} = c + \Delta/2$, which implies the existence of an even subsequence of time instants t_i $(= 2k_i)$ with

$$\tfrac{1}{2}E(|\, e_{t_i} \,| + |\, e_{t_i+1} \,|) \leqq \mathrm{K}. \tag{30}$$

Now define the truncating function $J_R(e)$ according to

$$J_R(e) = \begin{cases} 1 & |\, e \,| \leqq R \\ 1 - (|\, e \,| - R)/\delta & R < |\, e \,| < R + \delta. \\ 0 & R + \delta \leqq |\, e \,| \end{cases}$$

Then,

$$E(|\, e_{t_i} \,| + |\, e_{t_i+1} \,|) \geqq E\{|\, e_{t_i} \,|\, J_R(e_{t_i}) + |\, e_{t_i+1} \,|\, J_R(e_{t_i+1})\} \tag{31}$$

and, since the right-hand side is the expectation of a bounded continuous function,

$$E\{|\, e_{t_i} \,|\, J_R(e_{t_i}) + |e_{t_i+1} \,|\, J_R(e_{t_i+1})\} \to E_0 \,|\, e_j \,|\, J_R(e_j) + E_1 \,|\, e_{j+1} \,|\, J_R(e_{j+1}) \tag{32}$$

where j denotes an even time instant in steady-state operation and E_0 and E_1 denote the expectation under the distributions G_0 and G_1 respectively, or reversed, depending on the parity of x_1. Since eq. (32) holds for each positive R and δ, taking the limit as $\delta \to 0$ and $R \to \infty$ shows that

$$\tfrac{1}{2}(E_0 \,|\, e_i \,| + E_1 \,|\, e_{i+1} \,|) \leqq c + \Delta/2. \tag{33}$$

Thus, the mean-absolute quantizing noise averaged over two consecutive time instants has the bound $c + \Delta/2$ under the limiting distributions for ideal integration.

For leaky integration, the process x_k is bounded with probability one, so that in this case all moments of x_k converge to the corresponding

moment under the limiting distribution. Furthermore, in Appendix B, it was shown that the first absolute moment of u_k converges to the corresponding value under the limiting distribution. Together, this implies that $E \mid e_i \mid$ converges to $E_G \mid e_i \mid$. Hence, eq. (12) yields

$$E_G \mid e_i \mid \leqq c + \Delta/2\alpha. \tag{34}$$

This result could also have been obtained by the same argument used to derive eq. (33).

Note that the bounds, eqs. (33) and (34), are independent of the spectral density of the input process u_k and are therefore very crude bounds. An important feature is that the mean-absolute quantizing noise is shown to be finite under the stationary distributions. An immediate consequence is that the decoded process x_k has a finite first absolute moment for ideal as well as leaky integration. Since

$$E \mid x_i \mid = E \mid e_i + u_i \mid \leqq E \mid e_i \mid + E \mid u_i \mid ,$$

then

$$E_G \mid x_i \mid \leqq 2c + \Delta/2\alpha \tag{35}$$

for both leaky and ideal integration. (Set $\alpha = 1$ for ideal integration.) Possibly of interest also is that this bound may be used to obtain upper bounds on the tail probabilities of the decoded process by using the Chebyshev inequality.

As discussed in Section VI, ideal integration with initial values of x_1 of the form $m\Delta + \theta$ can be handled by replacing u_k by $u_k - \theta$, so that the bound, eq. (35), remains valid if c is replaced by $c + \mid \theta \mid$, and $\mid e_i \mid$ by $\mid e_i \mid - \mid \theta \mid$, which leads to the inequality, eq. (4).

IX. CONCLUSIONS

The results of this paper show that delta modulation indeed possesses the qualitative properties of convergence to a stationary distribution and boundedness of the quantizing noise. Perhaps of greatest interest is the fact that the results also hold for ideal integration, thus justifying the study of this idealized model to obtain an understanding of the usual physical situation of leaky integration.

The use of a Markovian model of the input process has been considered by several authors[8-11] as an approach to determine actual probability distributions for the steady state. The results of the paper show that the Markovian recursion, i.e., the usual Chapman–Kolmogorov equation, will, in fact, converge to the unique stationary distribution.

The technique used here for showing the existence of a stationary distribution (an invariant solution to the Chapman–Kolmogorov equation) extends the method used by this writer[12] for an adaptive filtering algorithm and the earlier results for Feller processes.[13,14]

APPENDIX A

Markovian Imbedding

Suppose initially that the process u_k has zero mean. Then u_k is generated by the recursion

$$u_{k+n} = \alpha_1 u_{k+n-1} + \cdots + \alpha_n u_k + \beta_1 w_{k+n-1} + \cdots + \beta_n w_k \quad (36)$$

with $\beta \neq 0$. This equation describes the operation of a stable shaping filter with transfer function

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=1}^{n} \beta_i z^{n-i}}{z^n - \sum_{i=1}^{n} \alpha_i z^{n-i}}$$

with $A(z)$ having all roots inside the unit circle, $|z| < 1$, and w_k is a white process with zero mean, finite variance and an everywhere-positive probability density function. The requirement that the autocovariance goes to zero for lags approaching infinity is satisfied by the fact that $B(z)$ is of lower degree than $A(z)$.

The state vector $\mathbf{d}_k$ is defined by

$$\begin{aligned} d_{1k} &= u_k - \mu \\ d_{i+1,k} &= d_{i,k+1} - b_i w_k \qquad i = 1, 2, \cdots, n-1 \end{aligned} \quad (37)$$

which when combined with eq. (36) leads to the state equations

$$\mathbf{d}_{k+1} = A\mathbf{d}_k + \mathbf{b}w_k \quad (38)$$

characterizing a vector Markov process. The matrix A is given by

$$A = \begin{bmatrix} 0 & 1 & 0 & & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & & \cdots & 1 \\ \alpha_n & \alpha_{n-1} & & & \cdots & \alpha_1 \end{bmatrix}$$

and the constant vector $\mathbf{b} = (b_1, b_2, \cdots, b_n)'$ is determined by solving the equations:

$$b_1 = \beta_1$$

$$b_i = \beta_i + \alpha_{i-1}b_1 + \cdots + \alpha_2 b_{i-2} + \alpha_1 b_{i-1} \quad \text{for} \quad i = 2, 3, \cdots, n.$$

The matrix A is stable since its eigenvalues are the roots of $A(z)$. This state representation is a standard one in the control literature. See for example, Ref. 15, p. 221.

The polynomials $B(z)$ and $A(z)$ may be assumed to have no common roots so that the shaping filter is intrinsically of order n. Then, the state generating eq. (38) is known to be completely controllable.* This means that, for any initial condition at time k, it is always possible to find values for $w_k, w_{k+1}, \cdots, w_{k+n-1}$ to produce any desired value of the state vector $\mathbf{d}_{k+n}$. It follows that since w_k has a positive density, the state vectors $\mathbf{d}_k$ have a positive probability of lying in any region of n-space with nonzero volume.

APPENDIX B

Convergence in Distribution of $\mathbf{d}_k$

The Markov process defined by eq. (38) can be iterated to obtain

$$\mathbf{d}_{k+1} = A^k d_1 + \sum_{i=1}^{k} A^{k-i} b w_i . \tag{39}$$

Since A has all eigenvalues of less than unit modulus, $A^k \to 0$ as $k \to \infty$, so that the first term on the right side of eq. (39) goes to zero with probability one. The second term has the same distribution as

$$v_k = \sum_{i=0}^{k-1} A^i b w_i$$

since the variates w_k are independent and identically distributed. But v_k is a martingale,[16] since

$$E\{v_{k+i} \mid v_k\} = v_k$$

and

$$E \mid\mid v_k \mid\mid \leqq \sum_{i=0}^{k-1} \lambda^i \mid\mid b \mid\mid E \mid w_i \mid < \frac{\mid\mid b \mid\mid}{1 - \lambda} E \mid w_k \mid$$

is finite, where λ denotes the Euclidean norm of A, $\lambda < 1$. Then by the

* See Theorem 7-8, p. 389 of Ref. 15.

martingale convergence theorem, v_k converges with probability one to a random variable v_∞. Hence, the probability distribution of the vector $\mathbf{d}_{k+1}$ converges to the distribution of v_∞, where

$$v_\infty = \sum_{i=0}^{\infty} A^i b w_i \,. \tag{40}$$

Since w_k is uncorrelated and has finite variance, it may be seen from eq. (40) that v_∞ also has finite variance. This, together with the convergence in distribution, implies (Ref. 6, p. 252) that the mean absolute value of each component of $\mathbf{d}_k$ converges to the mean absolute value of the corresponding component of v_∞. Consequently,

$$E \mid u_k \mid \rightarrow c$$

for any initial state $\mathbf{d}_1$, where c is the mean absolute value of the first component of v_∞.

APPENDIX C

Existence of a Stationary Distribution

Theorem: A Markov process has a stationary distribution if

(*i*) *for any initial state* $\mathbf{s}_1$ *the averaged distributions* G_k *are stochastically bounded, and*

(*ii*) *for any region* A, *let* D *be the set of points* $\mathbf{s}$ *at which the transition probability function* $p(\mathbf{s}, A)$ *is discontinuous and let* N_δ *be the set of all points whose distance from* D *is less than* δ; *then there is a function* $C(\delta)$ *independent of* $\mathbf{s}$ *with* $C(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ *and for all* $\mathbf{s}$,

$$p(\mathbf{s}, N_\delta) \leqq C(\delta).$$

Proof: Since the sequence of averaged distributions G_k is stochastically bounded, by the Helly selection theorem, Ref. 6, p. 267, there exists a subsequence G_{k_i} converging to a limiting distribution G. From the definition of G_k and T it follows that

$$TG_n = G_n + \frac{1}{n}(F_{n+1} - F_1)$$

so that

$$\mid TG_{n_i}(A) - G_{n_i}(A) \mid \leqq \frac{1}{n_i} \rightarrow 0$$

as $i \rightarrow \infty$ for any region A. Since G_{n_i} converges to G, it then follows that

$$TG_{n_i} \rightarrow G \qquad i \rightarrow \infty. \tag{41}$$

It remains to show that

$$TG_{n_i} \to TG \tag{42}$$

so that eqs. (41) and (42) will imply that $G = TG$, which is the desired result.

To prove (42), note that $G_{n_i} \to G$ implies that for any bounded continuous function φ of the state vector $(\mathbf{s})$,

$$E_i\varphi(\mathbf{s}) \to E_G\varphi(\mathbf{s}) \tag{43}$$

where E_i is the expectation under G_{n_i}. If $p(\mathbf{s}, A)$ were continuous in $\mathbf{s}$, eq. (42) would follow from eq. (43) by setting $\varphi(\mathbf{s}) = p(\mathbf{s}, A)$ and noting from eq. (9) that

$$E_i p(\mathbf{s}, A) = TG_{n_i}(A).$$

However, $p(\mathbf{s}, A)$ is not itself continuous and the following argument is needed to complete the proof.

Let $I_\delta(\mathbf{s})$ denote the function which is equal to 1 if $\mathbf{s}$ is not in N_δ, and for $\mathbf{s}$ in N_δ let $I_\delta(\mathbf{s})$ denote the distance of $\mathbf{s}$ from D. Then $I_\delta(s)p(\mathbf{s}, A)$ is bounded and continuous in $\mathbf{s}$ for any region A, and so

$$E_i\{I_\delta(\mathbf{s})p(\mathbf{s}, A)\} \to E_G\{I_\delta(\mathbf{s})p(\mathbf{s}, A)\}, \qquad i \to \infty.$$

But

$$E_i p(\mathbf{s}, A) - E_i\{I_\delta(\mathbf{s})p(\mathbf{s}, A)\} \leqq G_{k_i}(N_\delta) \tag{44}$$

and also

$$E_G p(s, A) - E_G\{I_\delta(\mathbf{s})p(\mathbf{s}, A)\} \leqq G(N_\delta). \tag{45}$$

But since

$$p(s, N_\delta) \leqq C(\delta)$$

by hypothesis, it follows by averaging over $\mathbf{s}$ that

$$F_k(N_\delta) \leqq C(\delta)$$

and therefore

$$G_{k_i}(N_\delta) \leqq C(\delta) \tag{46}$$

and, since $G_{k_i}(N_\delta) \to G(N_\delta)$, then

$$G(N_\delta) \leqq C(\delta). \tag{47}$$

Combining these results shows that

$$\limsup_{i\to\infty} | E_i p(s, A) - E_G p(s, A) | < 2C(\delta),$$

but since δ can be made arbitrarily small, it follows that

$$E_i p(\mathbf{s}, A) \to E_G p(s, A).$$

Hence, eq. (42) holds and the theorem is proved.

APPENDIX D

Existence of $C(\delta)$

From eq. (38), it follows that

$$u_{k+1} = (Ad_k)_1 + b_1 w_k$$

with $b_1 \neq 0$, so that the conditional distribution of u_{k+1} given $\mathbf{s}_k$ is continuously distributed because w_k is continuously distributed. Let $l = (Ad_k)_1$, and let

$$H(x) = P\{b_1 w_k < x\}.$$

Then

$$P\{u_{k+1} < x \mid \mathbf{s}_k\} = H(x - l) \tag{48}$$

which is a uniformly continuous function of x. Therefore, if

$$C(\delta) = \sup_x [H(x + \delta) - H(x - \delta)]$$

then

$$C(\delta) \to 0 \quad \text{as} \quad \delta \to 0.$$

But

$$P\{|\, u_k - x \,| < \delta \mid \mathbf{s}_k\} = H(x + l + \delta) - H(x + l - \delta) \leqq C(\delta)$$

using eq. (48), which proves the existence of a suitable function $C(\delta)$ independent of $\mathbf{s}_k$ and x.

APPENDIX E

Range of the Mapping $y = \pm 1 \pm \alpha \pm \alpha^2 \pm \cdots$

Theorem: For $\frac{1}{2} \leqq \alpha < 1$, the range of values taken on by the mapping

$$y = \sum_{i=0}^{\infty} \alpha^i b_i \tag{49}$$

for all binary sequences, $b_i = \pm 1$ each i, is the closed interval $|\, y \,| \leqq 1/(1 - \alpha)$.
Proof: For each y in the interval $|\, y \,| \leqq a$, with $a = 1/(1 - \alpha)$, generate a binary sequence $b_0, b_1, b_2, \cdots$ according to the algorithm below.

Let

$$p = \tfrac{1}{2}(y + a).$$

If $p \geqq 1$, let $f_0 = 1$ otherwise $f_0 = 0$. Let

$$s_n = \sum_{i=0}^{n} \alpha^i f_i \qquad n = 0, 1, 2, \cdots .$$

For $n = 1, 2, 3, \cdots$, if $p - s_n \geqq \alpha^{n+1}$ let $f_{n+1} = 1$, otherwise $f_{n+1} = 0$. Then

$$b_i = 2f_i - 1, \qquad i = 0, 1, 2, \cdots . \tag{50}$$

To prove that eq. (49) holds for the binary sequence generated in this manner, note first of all that for all $n \geqq 0$,

$$s_n \leqq s_{n+1}, \qquad \text{and} \qquad s_n \leqq p$$

so that $s_n \to s$ for some number s with $s \leqq p$. Suppose that $s < p$. Then there exists a largest integer m satisfying

$$p < s_{m-1} + \alpha^m. \tag{51}$$

Therefore, $f_m = 0$, and $f_{m+i} = 1$ for each $i > 0$. Consequently,

$$p > s = s_{m-1} + \alpha^{m+1} + \alpha^{m+2} + \cdots$$

so that

$$p > s_{m-1} + \alpha^{m+1}/(1 - \alpha). \tag{52}$$

But eqs. (51) and (52) imply that

$$\frac{\alpha^{m+1}}{1 - \alpha} < \alpha^m$$

so that

$$\alpha < \tfrac{1}{2},$$

which is a contradiction. Therefore,

$$p = \sum_{i=0}^{\infty} \alpha^i f_i$$

and so

$$y = \sum_{i=0}^{\infty} \alpha^i (2f_i - a) = \sum_{i=0}^{\infty} \alpha^i b_i ,$$

which proves the theorem.

REFERENCES

1. Slepian, D., "On Delta Modulation," unpublished work.
2. van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep., *8*, No. 5 (October 1953), pp. 367–385.
3. Goodman, D. J., "Delta Modulation Granular Quantizing Noise," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1197–1218.
4. Schindler, H. R., "Delta Modulation," IEEE Spectrum, *7*, No. 10 (October 1970). A recent survey with bibliography.
5. Whittle, P., *Prediction and Regulation by Linear Least-Square Methods*, Princeton: D. Van Nostrand Company, Inc., 1963, pp. 31–35.
6. Feller, W., *An Introduction to Probability Theory and Its Applications*, Second Edition, II, New York: John Wiley and Sons, Inc., 1966.
7. Doob, J. L., "Asymptotic Properties of Markoff Transition Probabilities," Trans. Amer. Math. Soc., *63*, (May 1948), pp. 393–421. See Theorem 5.
8. Fine, T. L., "The Response of a Particular Nonlinear System with Feedback to Each of Two Random Processes," IEEE Trans. Inf. Theory, 11–14, No. 2, (March 1968), pp. 255–268.
9. Stanley, T. P., and Aaron, M. R., unpublished work.
10. Protonotarios, E. N., "Application of the Fokker–Planck–Kolmogorov Equation to the Analysis of Differential Pulse Code Modulation Systems," J. Franklin Institute, *289*, (January 1970), pp. 31–45.
11. Mills, T. K., and O'Neal, J. B., Jr., "Quantizing Noise Calculations in Delta Modulation," Conference Record, IEEE Int. Symp. Commun., Montreal, 1971, pp. 1.1–1.4.
12. Gersho, A., "Adaptive Filtering with Binary Reinforcement," unpublished work.
13. Foguel, S. R., "Existence of Invariant Measures for Markov Processes; II," Proc. Amer. Math. Soc., *17*, 1966, pp. 387–389.
14. Beneš, V. E., "Finite Regular Invariant Measures for Feller Processes," J. Appl. Prob., *5*, 1968, pp. 203–209.
15. Ogata, K., *State Space Analysis of Control Systems*, Englewood Cliffs: Prentice-Hall, Inc., 1967.
16. Doob, J. L., *Stochastic Processes*, New York: John Wiley and Sons. Inc., 1953, Chapter 7.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Perspective Drawing of Surfaces With Hidden Line Elimination

By N. Y. GRAHAM

(Manuscript received February 18, 1972)

An efficient computer algorithm is described for the perspective drawing of a wide class of surfaces. The class includes surfaces corresponding to single-valued, continuous functions which are defined over rectangular domains. The algorithm automatically computes and eliminates "hidden lines." The number of computations in the algorithm grows linearly with the number of sample points on the surface to be drawn. An analysis of the algorithm is presented, and extensions to certain multi-valued functions are indicated. The algorithm is implemented and tested on two different computers: a large central computer with hard-copy capability, and a small laboratory computer affording interactive use. Running times are found to be exceedingly efficient on both machines. Interactive implementation of the algorithm, with on-line scope display and view-point control, enables effective and rapid examination of a surface from many perspectives.

I. INTRODUCTION

The general problem of efficiently and meaningfully displaying three-dimensional objects in two dimensions is central to computer graphics. In particular, the "hidden line problem" of drawing objects in perspective while eliminating line segments not visible from the viewing point, is one of long standing. In recent years various algorithms dealing with this problem have appeared in literature.[1-7] Some of these algorithms entail prohibitively long computation time (relative to the number of data points) or have large storage requirements; some have both of these undesirable characteristics.

In this paper a detailed description will be given of a highly efficient algorithm for the perspective drawing of an arbitrary surface which corresponds to a single-valued continuous function defined over a rectangular domain, with elimination of hidden lines. The vantage point from which the surface is viewed may be any point not on the

surface. Some generalizations of the algorithm to nonrectangular domains and to multiple-valued functions will be indicated.

This algorithm has been implemented on both a large and a small computer (Honeywell 6070, DDP 516) with considerably shorter computation time than that based upon previous algorithms for drawing similar surfaces. A very useful feature of the implementation on the DDP 516 (a 16K laboratory machine with 16-bit word length) is that the program may be run interactively, allowing the user to select varying vantage points and rapidly display different views of a given surface.

II. STATEMENT OF THE PROBLEM

Given a three-dimensional surface S defined over a rectangular domain R, and a vantage point V not on S (Fig. 1), we may assume that R is centered about the origin of the x, y-plane and is oriented so that its sides are parallel to the x-axis and y-axis, respectively. (This may be accomplished without loss of generality by a suitable translation and rotation of R, together with V). The plane P, containing the perspective image S' of S, is chosen to be perpendicular to the line joining V and the origin. A rectangular coordinate system with x', y'-axes is chosen on P which preserves the original vertical direction

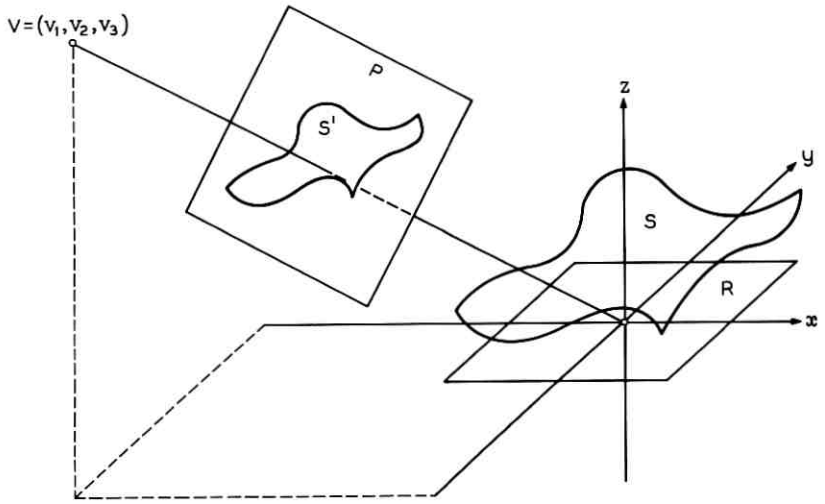

Fig. 1—Perspective projection of surface S on plane P.

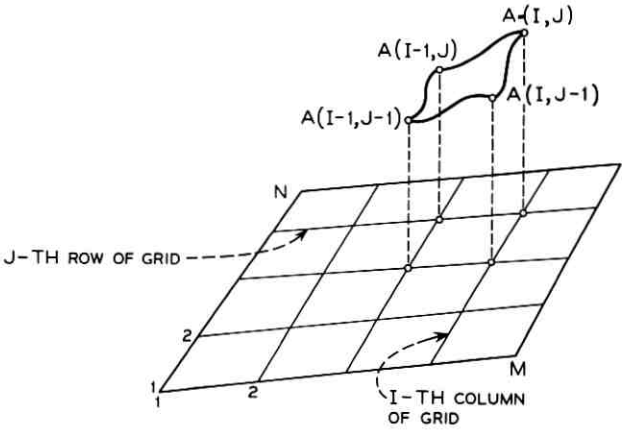

Fig. 2—Surface patch of one grid subrectangle.

of S. That is, the y'-axis on P must be the projection of the original z-axis. This latter choice is necessary, as we shall see later. (Relative to the rectangular system on P, the perspective image on P of each point of S has a well-defined pair of coordinates. The evaluation of these coordinates will be given in the Appendix.)

The surface S will first be quantized to an $M \times N$ rectangular grid of points on the domain R. Figure 2 shows the grid lines on R, partitioning R into subrectangles. The part of S defined over one subrectangle is shown. It will be referred to as a "surface patch" of S. After quantization, only the four points of the patch defined at the four vertices of the subrectangle are known, and linear interpolation of the function will be assumed between adjacent grid points, that is, between adjacent points in the same row and adjacent points in the same column of the grid. (Thus, S need not be explicitly defined by a mathematical function; the data for S may be given by a rectangular array of points corresponding to a rectangular grid of points on the domain.) The behavior of S in the interior of each subrectangle of the grid will be ignored. Thus, each surface patch of S will be represented by its four linearly interpolated edges in three-dimensional space. The image of these edges on the projection plane is a four-sided polygon (Fig. 3). For each surface patch of S, only the visible line segments of its edges will be drawn.

Intuitively, the surface should be thought of as an opaque elastic membrane stretched over a rigid frame consisting of all the linearly

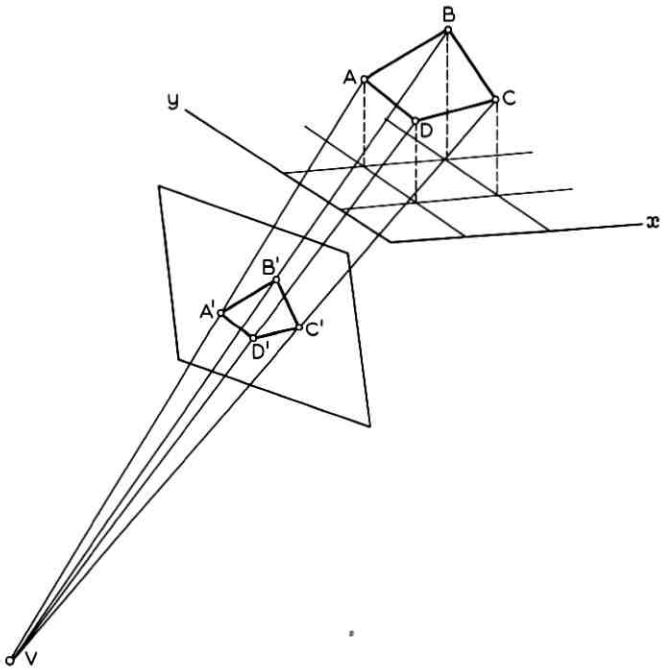

Fig. 3—Projection of linearly interpolated surface patch.

interpolated edges of the surface patches. Then the problem is to give a perspective drawing of the frame which eliminates line segments of the frame hidden by the opaque membrane from the vantage point. (Note that the vertical projection of the frame onto the x, y-plane coincides with the grid lines on R.)

III. BASIC IDEAS OF THE ALGORITHM

The algorithm for determining the visibility of any given edge is based upon two ideas. The first one consists of choosing a particular ordering of the surface patches of S so that no part of any patch occurring earlier in the ordering is obscured from the vantage point by any part of a patch occurring later. The surface is to be drawn according to such an ordering, one patch at a time. Then, a line segment on any edge of a patch under consideration is hidden from the vantage point if and only if its image on the projection plane lies inside a region of the plane already covered by the images of earlier patches.

This particular ordering of the surface patches has the further

property that, for a surface S which corresponds to a single-valued continuous function over a rectangular domain, the region of the projection plane being covered by the emerging image of S will grow contiguously as each patch in the ordering is drawn. And because of the preservation of the vertical direction of the surface, at each stage of the drawing of S, this "hidden" region of the plane can be precisely bounded between two piecewise linear continuous functions. The determination of the visibility of any edge of a surface patch is thus reduced to the problem of deciding which part of its image on the projection plane lies between these two functions. This delineation of the hidden region of the projection plane, after each patch is drawn, by means of two piecewise linear continuous functions, constitutes the second idea behind the algorithm.

Remark: If the y'-axis on the plane of projection is not chosen to be the perspective image of the original z-axis, then the image of the surface may be "tilted" so that it will be impossible to delineate its boundary by using only *two* functions.

IV. ORDERING OF THE SURFACE PATCHES

The fact that it is possible to order the surface patches of S so that earlier ones are not obscured by later ones depends on the following observation:

Suppose A and B are any two points of the surface such that A obscures B from V. Geometrically, this means that A, B, and V lie in a straight line, and the distance from A to V is less than the distance from B to V. Then it can be easily shown that the vertical projections A_0 , B_0 , V_0 of the points A, B, V, respectively, onto the x, y-plane satisfy this relationship also. That is, they are collinear, and A_0 is closer than B_0 to V_0 .

This observation is the key to the ordering of the surface patches. First, consider the case in which the vertical projection $V_0 = (v_1 , v_2 , 0)$ of the vantage point onto the x, y-plane is southwest of the domain, as shown in Fig. 4. [That is, $v_1 \leqq x_1$ and $v_2 \leqq y_1$, where $(x_1 , y_1 , 0)$ is the lower left-hand corner of the domain.] An immediate consequence of the above observation is that if $A = (a_1 , a_2 , a_3)$ obscures $B = (b_1 , b_2 , b_3)$ from V, then $a_1 < b_1$ and $a_2 < b_2$; i.e., A_0 must be southwest of B_0 (given the southwest location of V_0). This suggests the following rule for determining the relative order of points of S, given the southwest location of V_0 : Let A and B be any two points of S. If one of them, say A, is southwest of the other (more precisely, if A_0 is south-

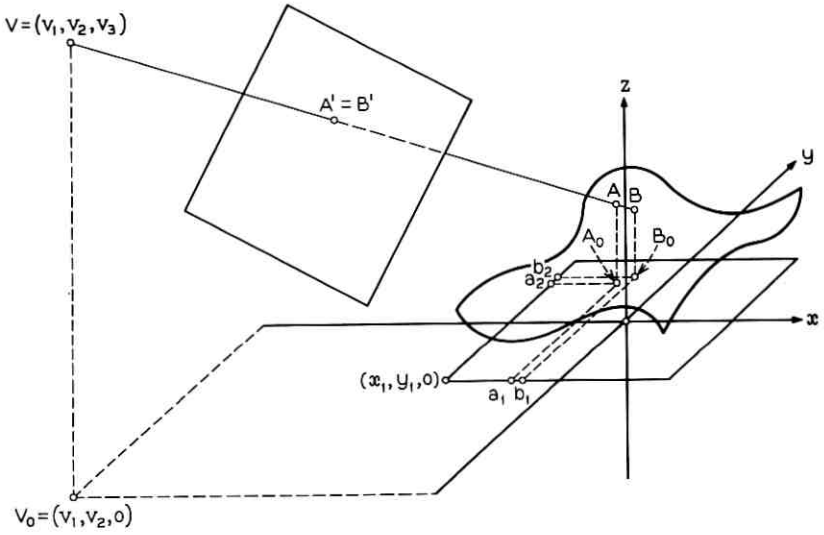

Fig. 4—Points of the surface with A obscuring B.

west of B_0), then A must precede B. Otherwise, A and B may be ordered independently of each other.

There are various ways to order the surface patches of S so that this rule for the relative order of points is satisfied, and hence, so that earlier patches are not obscured by later patches. Figure 5 illustrates one possible ordering of the subrectangles of R so that the induced

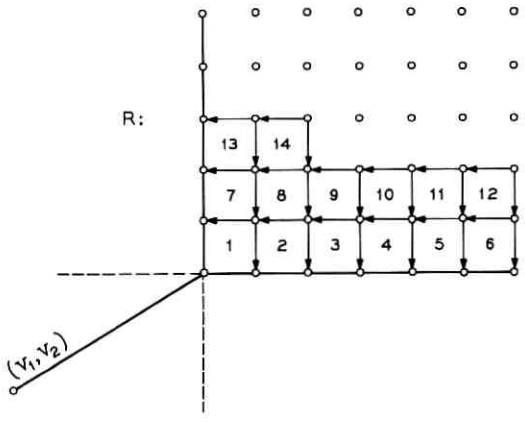

Fig. 5—One possible ordering of subrectangles.

ordering of the patches of S has this property; i.e., they are ordered by rows, from left to right in each row, beginning with the front row (relative to the southwest location of V_0). It is obvious that, under this ordering of the patches, no point of an earlier patch will be obscured by a point of a later patch, since no point of a later patch is southwest of any point of an earlier patch.

For other "corner" locations of V_0 (i.e., northwest, northeast, and southeast of the domain) analogous orderings of the surface patches may be made. Instead of discussing suitable orderings of the surface patches for other locations of V_0, we shall proceed to describe the algorithm in detail for the special case in which V_0 is southwest of the domain, and then show how the other cases may be reduced to cases involving only "corner" locations of V_0.

V. THE ALGORITHM IN DETAIL

Let S be a surface defined over a rectangular domain R of the x, y-plane. Assume S is to be viewed in perspective from a vantage point V whose vertical projection V_0 onto the x, y-plane is southwest of R. Given an $M \times N$ rectangular grid of points on R, for $I = 1, 2, \cdots, M$ and $J = 1, 2, \cdots, N$, let $A(I, J)$ denote the "sample point" on S defined at the grid point on R located at the Ith column and Jth row of the grid. For $I = 2, \cdots, M$ and $J = 2, \cdots, N$, let $S(I, J)$ denote the surface patch with vertices $A(I, J)$, $A(I, J - 1)$, $A(I - 1, J)$, $A(I - 1, J - 1)$. (See Fig. 2.) By above considerations, the patches may be ordered as follows, and drawn according to this ordering, one at a time:

$$S(2, 2), S(3, 2), \cdots, S(M, 2),$$
$$S(2, 3), S(3, 3), \cdots, S(M, 3),$$
$$\vdots$$
$$S(2, N), S(3, N), \cdots, S(M, N).$$

Now, consider the sequence of points $A(1, J)$, $J = 1, \cdots, N$, defined along the first column of grid points. The line segments joining these points in consecutive order will be called the "leading left edge" of S. Similarly, the line segments joining the sequence of points $A(I, 1)$, $I = 1, \cdots, M$ in consecutive order will be called the "leading front edge" of S. As a consequence of the observation made at the beginning of Section IV, both leading edges are entirely visible from V. Thus, we may begin by drawing all the line segments of the leading edges.

Then the drawing of each successive patch $S(I, J)$ in the ordering may be completed by adding the visible segments of just two of its four edges, namely its "back edge," joining $A(I, J)$ and $A(I - 1, J)$, and its "right edge," joining $A(I, J)$ and $A(I, J - 1)$, since the other two edges will have already been considered in connection with earlier patches or with one of the leading edges.

Figure 6a gives an example of a surface defined over a 5×3 grid of points. The 5×3 array of sample points on the surface gives rise to two rows of patches, four in the front row and four in the back row. Figure 6b shows the conjunction of the leading left edge and the leading front edge of this same surface.

Figures 7a through 7h show successive patches being added to the drawing of the surface. With each patch that is being added, L_1 denotes its "back" edge, and L_2 denotes its "right" edge. Dotted lines indicate hidden segments which are not drawn.

As indicated earlier, the determination of the visibility of each new edge under consideration will involve two piecewise linear continuous functions. Intuitively, one of these functions, Max, will be used to define the path of the upper perimeter of the region of the projection plane thus far covered, and the other function, Min, will be used to define the path of the lower perimeter. At this point, it would be useful to identify the grid of the plotting device with a rectangle in the projection plane which circumscribes the perspective image of the surface. (This corresponds to a scaling and translation of the coordinates of the image points so that they fall within the dimensions of the plotting grid.) Then, in order to optimize the precision in delineating the hidden region of the projection plane, or equivalently, the hidden region of the plotting grid, the number of breakpoints of each of the functions Max and Min is chosen to coincide with the number of horizontal units of the plotting grid.

Before plotting begins, we artifically set Max (k) = Min (k) = -1,

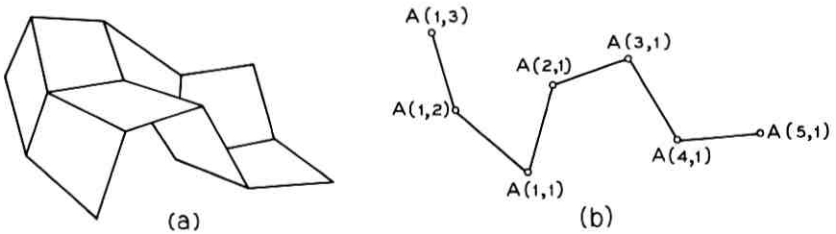

Fig. 6—Surface and its leading left and front edges.

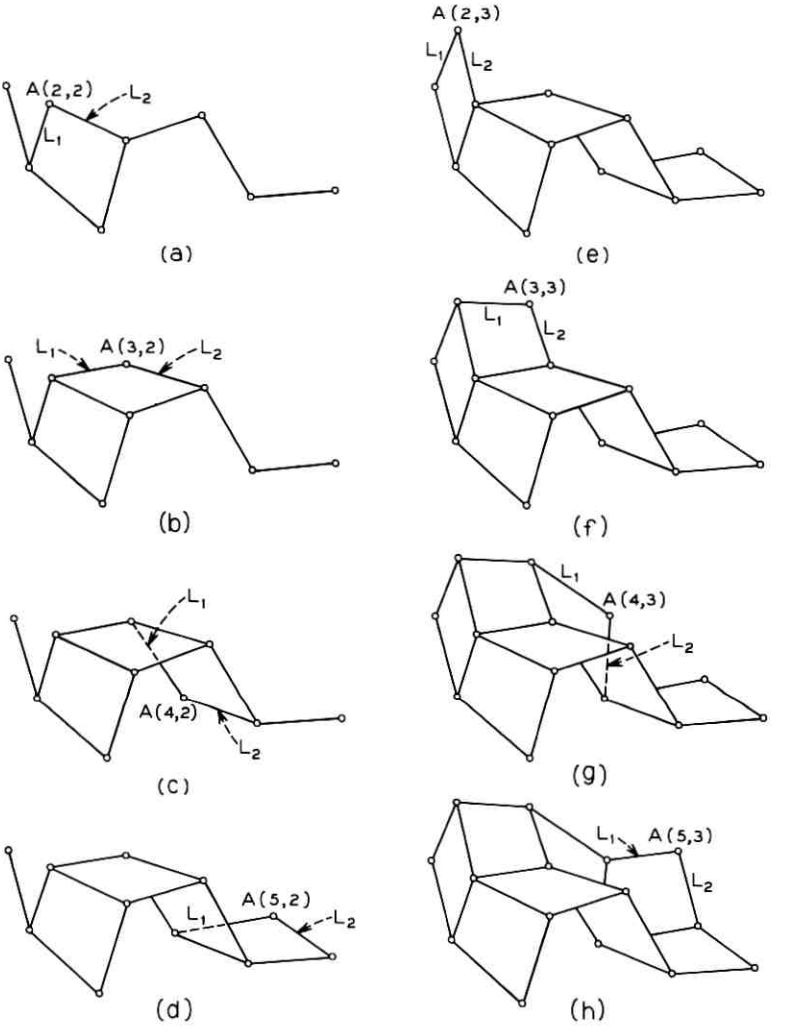

Fig. 7—Successive patches of surface being drawn.

for $k = 1, 2, \cdots, K_x$, where K_x is the horizontal dimension of the plotting grid. Then, after each line segment is drawn, Max or Min (possibly both) will be modified to reflect the change in the shape of the emerging image of S. We begin with the line segments on the leading edges of S. Let A and B be any two consecutive "sample points" of the leading left edge or of the leading front edge, that is, any two

consecutive points in the following sequence (see Fig. 2):

$$A(1, N), A(1, N-1), \cdots, A(1, 1), A(2, 1), \cdots, A(M, 1).$$

Let A_x , A_y be the (scaled and translated) plotting coordinates of A, and B_x , B_y the plotting coordinates of B. First, the line segment joining the points (A_x , A_y) and (B_x , B_y) is drawn. Then both Max and Min are redefined between A_x and B_x by setting

$$\text{Max}\,(A_x) = \text{Min}\,(A_x) = A_y$$

$$\text{Max}\,(B_x) = \text{Min}\,(B_x) = B_y$$

and linearly interpolating both Max and Min between A_x and B_x . (In Fig. 8, K_x and K_y refer to the dimensions of the plotting grid, in the horizontal and vertical directions, respectively.)

After the above procedure is carried out for every pair of consecutive points in the sequence, the leading edges will be completely drawn, and their path will be defined by Max and Min between the two endpoints. Note that the region bounded between Max and Min has area zero, reflecting the fact that no patch has yet been drawn.

We now proceed to draw the surface patches. Each $S(I, J)$ in the ordering, beginning with $S(2, 2)$ will be uniformly processed as follows: Let L_1 denote the edge between $A(I, J)$ and $A(I-1, J)$, L_2 denote the edge between $A(I, J)$ and $A(I, J-1)$. For each of these edges, we must first determine its visible segments, then draw only those segments, and modify Max or Min to reflect the change in the shape of the hidden region (the entire edge may be hidden, of course).

First, let $L = L_1$ and $A = A(I, J)$, $B = A(I-1, J)$, the endpoints of L_1 . The following criterion is used to determine the visibility of any point C on L: If C_x , C_y are the plotting coordinates of C, then

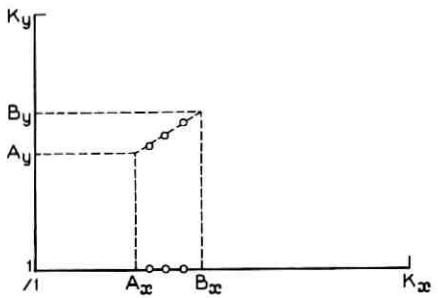

Fig. 8—Linear interpolation of Max and Min.

C is visible if and only if $C_y \geqq \text{Max}\ (C_x)$ or $C_y \leqq \text{Min}\ (C_x)$. This visibility criterion is applied to each of the endpoints A and B. (Let A_x , A_y be the plotting coordinates of A; B_x , B_y those of B.) There are four possibilities:

(*i*) A, B are both visible. Then all of L is assumed to be visible. (This assumption is made to reduce computation time. It may happen that some segments of hidden lines will be incorrectly drawn if the surface has sharp "spikes" in the foreground. This problem can be circumvented by starting with a finer grid on R.) The line segment joining (A_x , A_y) and (B_x , B_y) is drawn. If $A_y \geqq \text{Max}\ (A_x)$, Max is linearly interpolated between A_x and B_x by setting $\text{Max}\ (A_x) = A_y$, $\text{Max}\ (B_x) = B_y$. Otherwise, we must have $A_y \leqq \text{Min}\ (A_x)$, and Min is similarly modified between A_x and B_x , with $\text{Min}\ (A_x) = A_y$, $\text{Min}\ (B_x) = B_y$.

(*ii*) A and B are both hidden. Then for the same reasons as in case (*i*), we assume all of L is hidden and no line segment is drawn. Max and Min are left unchanged, and we proceed to the next edge.

(*iii*) A is hidden and B is visible. A search is made along the line joining A and B, starting at A, for the first visible point C (discrete steps are taken in the horizontal direction of the plotting grid). The segment joining C to B is assumed to be entirely visible and is drawn. If $C_y \geqq \text{Max}\ (C_x)$, Max is linearly modified between C_x and B_x with $\text{Max}\ (C_x) = C_y$, $\text{Max}\ (B_x) = B_y$. Otherwise, Min is modified between C_x and B_x .

(*iv*) A is visible and B is hidden. A search is made along the line joining A and B for the first visible point C, starting from the hidden point B. The rest is analogous to case (*iii*).

For each surface patch $S(I, J)$, this procedure is carried out with $L = L_1$ and $B = A(I - 1, J)$; then the procedure is repeated with $L = L_2$ and $B = A(I, J - 1)$. This completes the description of the algorithm for the special case in which V_0 , the vertical projection of the vantage point onto the x, y-plane, is southwest of the domain R.

If V_0 is in any of the other "corner" regions of the x, y-plane (indicated by NW, NE, and SE in Fig. 9a), we may rotate the data defining the surface (either the mathematical function or the rectangular array of sample points) by 90, 180, or 270 degrees, respectively, and apply the algorithm as described for V_0 in the SW region of the x, y-plane. Another approach is to change the ordering of the surface patches and to define new leading edges to suit the other locations of V_0 .

If V_0 is in any of the regions directly west, north, east, or south of R

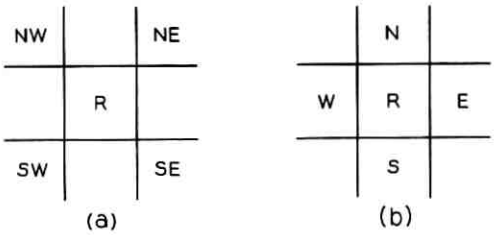

Fig. 9—Locations of vertical projection of vantage point.

(Fig. 9b), we may partition R into two subrectangles R_1 and R_2 (Fig. 10a), and apply the algorithm to each of R_1 , R_2 , since V_0 is in a "corner" location relative to each subrectangle. The two parts of the surface, corresponding to R_1 and R_2 , will have to be "pieced together" along the line of partition, and this may be done without "flaws" by augmenting the sample points of the surface with the set of points defined at the intersection of the partition line with the grid lines on R.

Finally, if V_0 lies inside the domain R, we may partition R into four subrectangles (Fig. 10b) and apply the algorithm to each of R_1 , R_2 , R_3 , R_4 , with V_0 in a "corner" location.

Remark: Note that under the assumption that the plane of projection P be perpendicular to the line joining V and the origin, if V_0 coincides with the origin, then the projection of the z-axis from V onto P is a point. In this case, the y'-axis on P must be chosen to be the image of some other line. See the Appendix.

Figures 11, 12, and 13 are sample drawings generated on the Honeywell 6070 computer with the Stromberg–Carlson 4060 microfilm plotter. Each surface was evaluated at a 64 × 64 grid of points, for a total of 4096 data points. The average run time was 9 seconds for each surface. (Part of the 9 seconds was used in computing the function values and

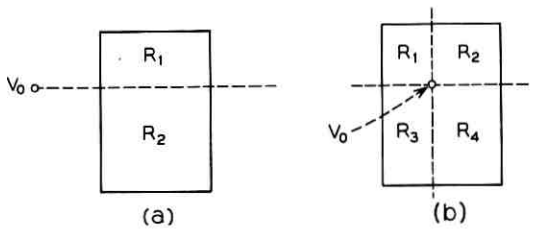

Fig. 10—Partitioning of domain for non-corner locations of V_0.

plotting coordinates at each point of the 64 × 64 grid. In fact, these computations were performed twice, once for determining the scale factors and again for the actual plotting of the surface.) Total storage for generating microfilm output of each surface was 16K.

It is clear that using this algorithm, computation time should vary linearly with the number of data points, since each data point is processed by evaluating its plotting coordinates IX, IY, and then comparing IY with just two values, Max (IX) and Min (IX).

Economically generated movies have also been produced on both the Honeywell 6070 and the DDP 516 computers. These movies show surfaces in rotation and surfaces undergoing continuous change in shape. The average run time on the Honeywell 6070 for making such a movie with a 32 × 32 grid and viewed from 240 distinct vantage points was 7.4 minutes, or about 1.85 seconds per frame. The corresponding run time on the DDP 516 was 7.5 seconds per frame. (For these movies, the scale factors were precomputed so that function values and plotting coordinates were evaluated only once for each frame of the movie.)

For a rough comparison of computation times for hidden-line elimination drawings based upon other algorithms (and implemented on other

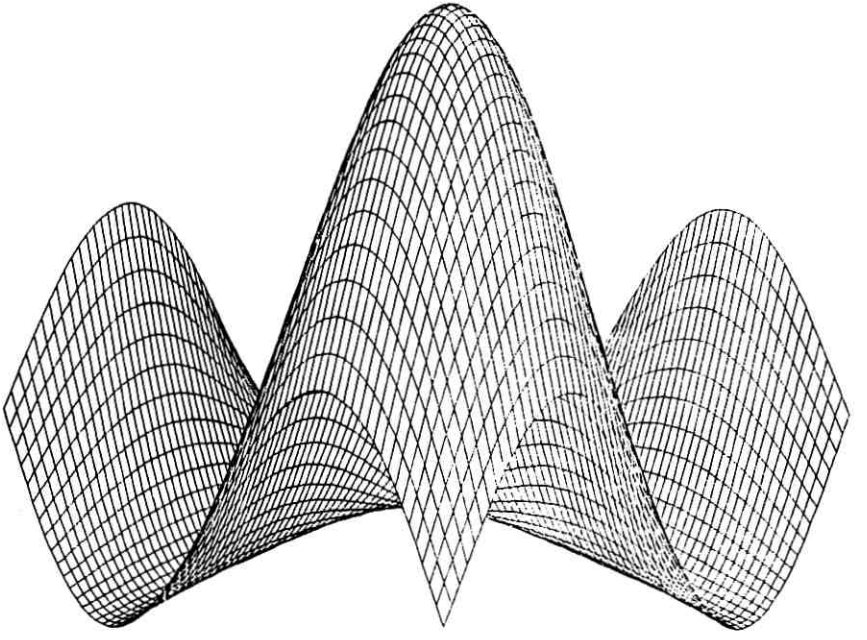

Fig. 11—$f(x, y) = (\sin(x) - 1)(\sin(y) - 1)$, $-\pi \leqq x, y \leqq \pi$.

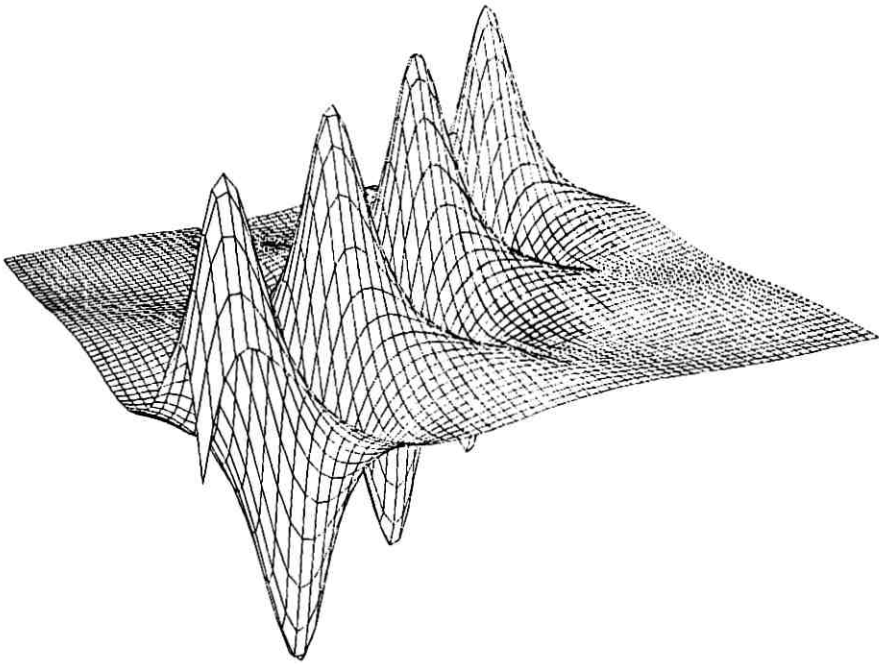

Fig. 12—$f(x, y) = \sin(x + y)/(1 + (x - y)^2)$, $-2\pi \leqq x, y \leqq 2\pi$.

computers) see Refs. 1 through 7. In particular, see Refs. 2 and 3 for a comparison with two other algorithms which deal with single-valued functions defined over a rectangular domain. Using the algorithm described in Ref. 2, computation time increases quadratically with the number of data points, since the visibility of each data point is determined by an exhaustive comparison with every other data point. Although computation time increases linearly using the algorithm of Ref. 3, the storage required is large in order to obtain high-quality drawings. This is because the "plotting page" is subdivided into rectangles, and, of all possible line segments of the surface whose images lie in any given rectangle, only the one closest to the vantage point is drawn in that rectangle. Thus, in order to avoid a "sketchy" drawing, a fine subdivision of the "plotting page" is necessary and a large array is required to store the information in each rectangle of the subdivision. (The algorithm of Ref. 3, however, is applicable to a larger class of surfaces than the algorithm described in this paper. For example, it can be used to draw surfaces corresponding to functions which are not single-valued, e.g., intersecting cylinders.)

VI. GENERALIZATIONS

6.1 *Convex Domains*

The algorithm may be generalized to an arbitrary surface S defined by a single-valued continuous function over a *convex* domain R, by artificially extending the defining function f to a rectangular domain R' containing R. The ideas of the algorithm are then applied to the surface S' defined by the extended function f' over R', with the following modifications: First, the sample points along the leading edges of S' must be replaced by a subset of points defined at the intersection of the boundary of R with the grid lines on R', and special consideration must be given to drawing the lines connecting these points. Second, a test must be made of sample points of S' to determine if they belong to S, so that only line segments that are part of S are drawn. One way to accomplish this is by setting $f'(x, y) = z_0$ for all (x, y) in $R' - R$, where z_0 is a number outside the range of values of f on R. (Convexity of the domain is necessary in order to bound the emerging images at each stage of the drawing between *two* functions.)

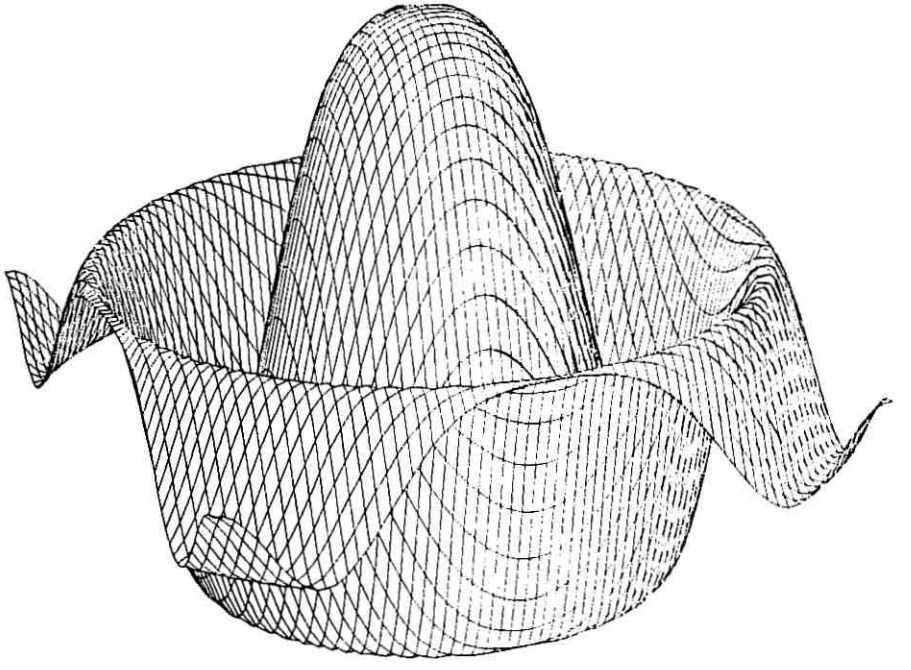

Fig. 13—$f(x, y) = \cos(x^2 + y^2)/e^{0.2(x^2+y^2)}$ $-2.5 \leqq x, y \leqq 2.5$.

6.2 *Multi-Valued Functions*

The algorithm may also be extended to certain multi-valued functions. For example, if the surface is a sphere, it is first partitioned by a horizontal plane into two single-valued pieces, and if the vantage point is above the horizontal plane, the algorithm is first applied to the upper hemisphere, and then, without reinitializing Max or Min, the algorithm is applied to the lower hemisphere.

6.3 *Bivariate Histograms*

With minor modifications the algorithm may be adapted to the perspective drawing of bivariate histograms, with elimination of hidden lines. Figures 14a through 14d show four different perspectives of a bivariate histogram generated by the same data. The computation time on the Honeywell 6070 was 1.6 seconds for each of the four drawings.

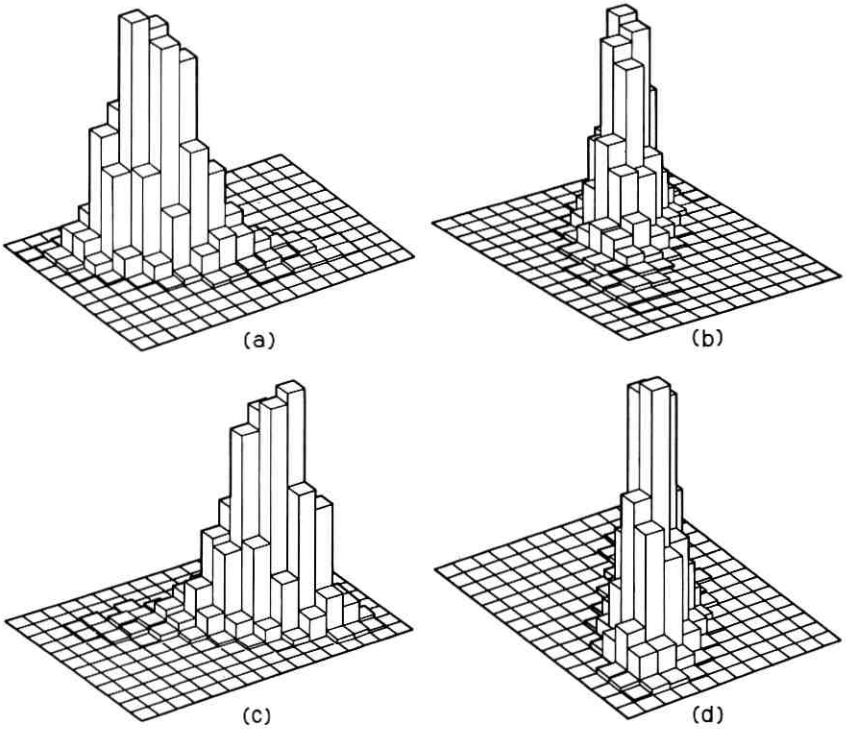

Fig. 14—Bivariate histograms: (a) V_0 in SW corner. (b) V_0 in SE corner (c) V_0. in NE corner. (d) V_0 in NW corner.

APPENDIX

Evaluation of Rectangular Coordinates of Image Points

Let S be a surface defined over a domain centered at the origin of the x, y-plane, and let $V = (v_1 , v_2 , v_3)$ be a point not on S. In order to simplify the calculation of the coordinates of image points and hence reduce computation time, a special plane P^*, normal to $\bar{V}$ and passing through the origin, will be used as the plane of projection. (In Fig. 1, a plane P parallel to P^* is shown. Note that the image of S on P differs from the image of S on P^* only by a scale factor.)

Let $A = (a_1 , a_2 , a_3)$ be any point on S, and let $B = (b_1 , b_2 , b_3)$ denote its image point on P^*. The plotting coordinates of A depend on the two-dimensional coordinates (u, v) of B with respect to an orthogonal system on P^*. If $\bar{X} = (x_1 , x_2 , x_3)$ and $\bar{Y} = (y_1 , y_2 , y_3)$ are the positive unit vectors of the coordinate system on P^*, then the scalars u and v must satisfy the following relationship:

$$u\bar{X} + v\bar{Y} = \bar{B}. \tag{1}$$

Since B is collinear with the points A and V, then

$$\bar{B} = \bar{V} + r(\bar{A} - \bar{V}), \quad \text{for some scalar } r. \tag{2}$$

Thus, eq. (1) may be replaced by

$$u\bar{X} + v\bar{Y} = \bar{V} + r\bar{A} - r\bar{V}. \tag{3}$$

Taking the dot product of both sides of eq. (3) with $\bar{X}$, we have

$$u\bar{X}\cdot\bar{X} + v\bar{Y}\cdot\bar{X} = \bar{V}\cdot\bar{X} + r\bar{A}\cdot\bar{X} - r\bar{V}\cdot\bar{X}.$$

Since $\bar{X}\cdot\bar{X} = 1$ and $\bar{Y}\cdot\bar{X} = \bar{V}\cdot\bar{X} = 0$ ($\bar{V}\cdot\bar{X} = 0$ because $\bar{X}$ on P^* and P^* normal to $\bar{V}$), we have

$$u = r\bar{A}\cdot\bar{X}. \tag{4}$$

Similarly, taking the dot product of both sides of eq. (3) with $\bar{Y}$, we have

$$v = r\bar{A}\cdot\bar{Y}. \tag{5}$$

The scalar r is determined by taking the dot product of both sides of eq. (2) with $\bar{V}$. Since $\bar{B}$ lies on P^*, then $\bar{B}\cdot\bar{V} = 0$, and we have

$$r = \bar{V}\cdot\bar{V}/\bar{V}\cdot(\bar{V} - \bar{A}). \tag{6}$$

It remains to determine the coordinates of the unit vectors $\bar{X}$, $\bar{Y}$.

As mentioned in Section II, the y'-axis on P^* must be the perspective

image of the original z-axis. Note, however, that if V is on the z-axis, then the perspective image from V of the z-axis is a single point, which of course cannot be used as an axis line. In this case, we must compromise by either moving the vantage point slightly off the z-axis or choosing the y'-axis on P^* to be the image of a slightly "askew" z-axis. For the rest of this discussion we assume that V is not on the z-axis.

Let $Z = (0, 0, 1)$. Then $\bar{V}$, $\bar{Y}$, $\bar{Z}$ lie on the same plane, because $\bar{Y}$ is the image from V of some scalar multiple of $\bar{Z}$. Thus $\bar{Z}$ is a linear combination of $\bar{V}$ and $\bar{Y}$. Since $\bar{X} \cdot \bar{V} = 0$ and $\bar{X} \cdot \bar{Y} = 0$, then $\bar{X} \cdot \bar{Z} = 0$. The normalization of any vector $\bar{X}'$ satisfying the equations $\bar{X}' \cdot \bar{V} = 0$ and $\bar{X}' \cdot \bar{Z} = 0$ will be a unit vector on the x'-axis of P^*. $\bar{X}' = \overline{(-v_2, v_1, 0)}$ satisfies these two equations. Let $d_1 = \sqrt{v_1^2 + v_2^2}$. Then

$$\bar{X} = \overline{(-v_2/d_1 , v_1/d_1 , 0)}$$

is the positive unit vector on the x'-axis of P^*. The coordinates of $\bar{Y}$ are similarly determined from the equations $\bar{Y} \cdot \bar{X} = 0$ and $\bar{Y} \cdot \bar{V} = 0$. Let $d = \sqrt{v_1^2 + v_2^2 + v_3^2}$. Then

$$\bar{Y} = \overline{(-v_1 v_3 , -v_2 v_3 , d_1^2)/dd_1}$$

is the positive unit vector on the y'-axis of P^*. Substituting these values of $\bar{X}$ and $\bar{Y}$ back into eqs. (4) and (5), we have

$$u = r(a_2 v_1 - a_1 v_2)/d_1 \tag{7}$$

$$v = r(-a_1 v_1 v_3 - a_2 v_2 v_3 + a_3 d_1^2)/(dd_1), \tag{8}$$

where $r = d^2/(d^2 - (a_1 v_1 + a_2 v_2 + a_3 v_3))$, as defined in eq. (6).

The above expression for v is algebraically equivalent to the following one, which involves half as many arithmetic operations:

$$v = (v_3 + r(a_3 - v_3)) \cdot d/d_1. \tag{9}$$

Thus, given a point A on S, the rectangular coordinates (u, v) of its image point are evaluated according to eqs. (7) and (9).

REFERENCES

1. Roberts, L. G., "Machine Perception of Three-Dimensional Solids," Technical Rep. No. 315, Lincoln Laboratory, M.I.T., (May 1963).
2. Kubert, B., Szabo, J., and Giulieri, S., "The Perspective Representation of Functions of Two Variables," J. Assn. for Computing Machinery, *15*, No. 2, (April 1968).
3. Kubert, Bruce R., "A Computer Method for Perspective Representation of Curves and Surfaces," Technical Rep. No. TR-0200-2, Aerospace Corporation, (December 1968).

4. Galimberti, R., and Montanari, U., "An Algorithm for Hidden Line Elimination," Commun. Assn. for Computing Machinery, *12*, No. 4, (April 1969).
5. Loutrel, P., "A Solution to the Hidden-Line Problem for Computer-Drawn Polyhedra," IEEE Trans. Computers, *C-19*, No. 3, (March 1970).
6. Encarnacao, J. L., "A Survey of and New Solutions for the Hidden-Line Problem," Symp. on Interactive Computer Graphics, (October 1970).
7. Weiss, Ruth A., "BE VISION, A Package of IBM 7090 Fortran Programs to Draw Orthographic Views of Combinations of Plane and Quadric Surfaces," J. Assn. for Computing Machinery, *13*, No. 2, (April 1966).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Some Effects of Quantization and Adder Overflow on the Forced Response of Digital Filters

By A. N. WILLSON, JR.

(Manuscript received December 28, 1971)

The effects of quantization (i.e., roundoff, truncation, etc.) and adder overflow, which are present in any special-purpose computer type realization of a digital filter, cause an otherwise linear system to become quite nonlinear. Moreover, the presence of such nonlinearities can cause the system's response to differ drastically from the ideal response (that is, from the response of the linear model of the filter) even when the level of the filter's input signal is, in a certain reasonable sense, small, and when the quantization effects are made arbitrarily small.

In this paper we derive a criterion for the satisfactory behavior of second-order digital filters in the presence of such nonlinear effects. The criterion is shown to be sharp, in that we also present a procedure for constructing counterexamples which show that, for most filters which violate the criterion, the response to some "small" nonzero input signal is not always even asymptotically close to the ideal response.

I. INTRODUCTION

The effects of quantization (i.e., roundoff, truncation, etc.) and adder overflow are present in any special-purpose computer type realization of a digital filter. When taken into account, these effects cause an otherwise linear system to become quite nonlinear. To date, the analysis of limit cycle phenomena in such nonlinear digital filters has been concerned with the study of the zero-input response of second-order filters.[1-3] A more fundamental problem is that of determining whether or not a filter's response to a nonzero input (the forced response) is in some meaningful sense close to the ideal response. This problem seems to have been ignored.

If we consider input sequences, the levels of which are sufficiently small (in the sense that when the input sequence is applied to the linear

model of the filter, the response *eventually* lies within the open interval determined by the most positive and the most negative machine numbers), then it is tempting to conjecture, as if the system were linear, that when the filter's zero-input response can be made to admit only limit cycles of small amplitude by using sufficiently many bits in the representation of the data so that the quantization errors are made sufficiently small, then the deviation of the filter's forced response from the ideal can also be made small in the same manner. As will be shown by counterexamples, however, this conjecture is false. Thus, since the usual purpose of a digital filter is the processing of nonzero signals, a question of major importance becomes: How can it be determined that, in the presence of quantization and adder overflow, a digital filter's forced response will be satisfactory?

In this paper we analyze the forced response of second-order digital filters which employ a type of arithmetic that has been called *saturation arithmetic.*† The essential structure of a second-order digital filter is shown in Fig. 1 where, for given real numbers a, b the filter's output sequence‡ $v^{(k)}$, $k = 1, 2, \cdots$, is uniquely determined by the input sequence $u^{(k)}$, $k = 1, 2, \cdots$, and by $v^{(-1)}$, $v^{(0)}$, the initial values of the filter's state variables. We develop a criterion by which satisfactory behavior of the filter can be determined. The criterion is shown to be sharp, in the sense that our counterexamples show that for most filters which violate the criterion, the forced response is not always close to the ideal response.

More precisely, we show that when the filter's coefficients a, b are determined by any point lying within the open crosshatched region of Fig. 2, and for any input sequence whose level is small (in the sense mentioned earlier), then the response of the nonlinear filter will be asymptotically close to the ideal response. On the other hand, we show that when the filter's coefficients are determined by any point lying within the shaded regions in the lower corners of the triangle of Fig. 2, and when certain very reasonable assumptions are satisfied concerning the nature of the quantization, then there exist input sequences the levels of which are also small, but for which the filter's response is not asymptotically close to the ideal response.

---

† The definition of this term is given in Section II.

‡ In many applications some linear combination of the quantities $v^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ is taken to be the filter's output at the kth time instant. This additional complication has no bearing on the matters considered here. For simplicity, therefore, we consider the sequence $v^{(k)}$ to be the filter's output.

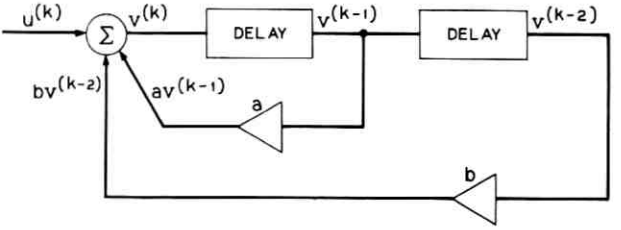

Fig. 1—Second-order digital filter.

II. SECOND-ORDER FILTERS

The usual method of designing digital filters[4] employs the interconnection of many second-order filters. The analysis and design of second-order digital filters is therefore a problem of considerable practical importance.

The behavior of the digital filter of Fig. 1 is characterized by the linear difference equation

$$w^{(k+1)} = Aw^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{1a}$$

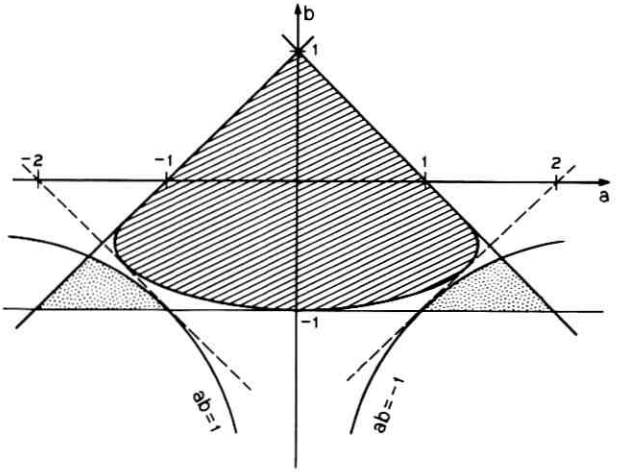

Fig. 2—Region determining filter coefficients for which the filter's forced response can be made asymptotically close to the ideal response (crosshatched region), and region determining coefficients for which the forced response will not always be close to the ideal response (shaded region).

where A denotes the 2×2 matrix

$$A = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix}, \tag{1b}$$

and $w^{(k)}$ is a two-dimensional vector (specifying the state of the system at the kth time instant) the second component of which, $w_2^{(k)}$, corresponds to the digital filter's output sequence $v^{(k)}$.

In any special-purpose computer type realization of the digital filter of Fig. 1 the ideal behavior specified by (1) can be only approximated. At each time instant, the output of the summation point can assume only one of a finite number of values. Therefore, the actual value of the summation point's output is given by an expression such as

$$v^{(k)} = f(av^{(k-1)} + bv^{(k-2)} + u^{(k)}) + e^{(k)},$$

where the function f accounts for adder overflow and the sequence $e^{(k)}$ accounts for the quantization error that is inherently present. The equality $f(\xi) = \xi$ is satisfied only in a certain neighborhood of the origin which we take to be the interval $-1 \leqq \xi \leqq 1$. We consider filters employing *saturation arithmetic*; that is, we define $f(\xi) = -1$ for $\xi < -1$ and $f(\xi) = 1$ for $\xi > 1$.

When the effects of quantization and adder overflow are taken into consideration, the digital filter of Fig. 1 is then characterized by the nonlinear difference equation

$$r^{(k+1)} = F\left(Ar^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix}\right) + \begin{bmatrix} 0 \\ e^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{2}$$

where the state of the system at the kth time instant is now specified by the two-dimensional vector $r^{(k)}$. The mapping F is defined by the relation

$$F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x \\ f(y) \end{bmatrix}. \tag{3}$$

Since the purpose of our study is to examine the effects of quantization and adder overflow on the forced response of digital filters, we are interested in comparing the solutions of (1) and (2) when the equations are given identical input sequences and identical initial conditions. We make the reasonable assumption that we are concerned only with digital filters whose linear model, i.e., eq. (1), is asymptotically stable. It is well known that (1) characterizes an asymptotically stable linear

system if and only if each eigenvalue of the matrix A has magnitude less than unity. The eigenvalues of A (the roots of the polynomial $\lambda^2 - a\lambda - b$) are known to have magnitude less than unity if and only if the coefficients a, b have values determined by points that lie within the large open triangular region shown in Fig. 2 (determined by the straight lines: $b \pm a = 1$, $b = -1$).

It is clear that so long as the filter's input sequence is such that the solution of the linear equation (1) is continually being driven into the region† $\| w^{(k)} \| > 1$, then there is little point in trying to compare the solutions of (1) and (2); it being clear at the start that at each such time instant, they will differ by at least the amount by which $\| w^{(k)} \|$ exceeds unity (plus or minus the quantization error $e^{(k)}$ which, presumably, will be small). At the other extreme, if it is known in advance that the initial conditions and the input sequence are (small enough in magnitude) such that the solution of the linear equation (1) is within the range $\| w^{(k)} \| \leqq 1 - \delta$, for some $\delta > 0$, and for all $k = 1, 2, \cdots$, then there is no problem. That is, it is clear at the outset (due to the assumption that the linear system is asymptotically stable) that the solutions of (1) and (2) will be made arbitrarily close for all such inputs, by simply causing the magnitude of the quantization error $e^{(k)}$ to be bounded by a sufficiently small number. In effect, the nonlinear function f is then not present; we are simply comparing the responses of the same stable linear system to two slightly different inputs.

The interesting question which we shall consider is the one which follows. Suppose we assume only that the filter's input is such that the ideal response, the solution of the linear system (1), *eventually* (i.e., for all k sufficiently large) satisfies $\| w^{(k)} \| \leqq 1 - \delta$, for some $\delta > 0$.‡ Then, when is the same thing (i.e., $\| r^{(k)} \| \leqq 1 - \delta$ for some $\delta > 0$, and all k sufficiently large) true for the solution of eq. (2)? Thus, we are interested in knowing when the gross effects of the nonlinearity are simply of a transient nature and hence, aside from such transient effects, when can the filter's response be made as close to the ideal as desired by simply causing the quantization error to be sufficiently small (i.e., by using a sufficient number of bits in the representation of the data). Unfortunately, as our counterexamples will show, it is *not* always the case

† For each $w = (w_1, w_2)^T$ we define $\|w\| = \max \{|w_1|, |w_2|\}$.

‡ The inequality $\|w^{(k)}\| \leqq 1$ might seem more reasonable here. The necessity to write $1 - \delta$ on the right-hand side is the small price that we must pay for the freedom to treat the quantization error in the relatively simple manner that we have chosen. By considering the quantization error at each step to be simply a "small" input $e^{(k)}$, we do not admit to the knowledge that, for example, in all sufficiently small neighborhoods of the points $\xi = \pm 1$, the quantization (be it roundoff, truncation, or whatever) will be done in such a manner that $|\xi + e^{(k)}| \leqq 1$.

that this will occur in the nonlinear system whenever the linear system's response satisfies $\| w^{(k)} \| \leqq 1 - \delta$ for some $\delta > 0$ and all sufficiently large k.

With our objective thus being to compare the asymptotic behavior of the solutions of (1) and (2), and since the linear system (1) is assumed to be asymptotically stable, it is clear that we may drop the requirement that the equations have the same initial conditions. This follows, of course, from the fact that the initial conditions of (1) do not affect the solution's asymptotic behavior.

By including the quantization effects in the linear model of the filter, the system is then described by the equation

$$s^{(k+1)} = As^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix} + \begin{bmatrix} 0 \\ e^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{4}$$

whose solution can be made arbitrarily close to the solution of (1) by simply requiring that all $| e^{(k)} |$ be sufficiently small. Let us assume, therefore, that the $| e^{(k)} |$ are at least small enough that there exists $\delta' > 0$ and a nonnegative integer K such that, for all nonnegative integers $k \geqq K$,

$$\| s^{(k+1)} \| + | e^{(k+1)} | \leqq 1 - \delta'. \tag{5}$$

Letting

$$z^{(k)} = r^{(k)} - s^{(k)}, \qquad k = 0, 1, 2, \cdots, \tag{6}$$

we find, from (2) and (4), that the sequence $z^{(k)}$ is determined by the equation

$$z^{(k+1)} = F\left(Az^{(k)} + \begin{bmatrix} 0 \\ v^{(k+1)} \end{bmatrix} \right) - \begin{bmatrix} 0 \\ v^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{7}$$

where $z^{(0)} = r^{(0)} - s^{(0)}$ and, for $k = 0, 1, 2, \cdots, v^{(k+1)} = s_2^{(k+1)} - e^{(k+1)}$ which, according to (5), with $\epsilon = 1 - \delta'$, satisfies

$$| v^{(k+1)} | \leqq \epsilon, \qquad \text{for } k \geqq K. \tag{8}$$

We take as our objective, therefore: To determine when, for any sequence $v^{(k+1)}$, $k = 0, 1, 2, \cdots$, satisfying (8) for some ϵ in the interval $0 \leqq \epsilon < 1$, and some nonnegative integer K, the solution of (7) satisfies $\lim_{k \to \infty} \| z^{(k)} \| = 0$.

We note at this point that our objective stated in the preceding paragraph is similar to the objective in Ref. 3 (see the paragraph immediately following eq. (8) of that paper) where the control of

limit cycles in the zero-input response of second-order digital filters is considered. The important difference between the two objectives is that here we must accommodate any value of ϵ in the interval [0, 1). In Ref. 3, however, it was only necessary to consider bounds on the sequence $\nu^{(k+1)}$ that were "sufficiently small". The consequences of this difference are great. It will be clear that a much more delicate analysis is required here than that in Ref. 3.

III. ANALYSIS OF THE FORCED RESPONSE

We now determine, in accordance with the objective explained in Section II, a criterion for the satisfactory behavior of the forced response of second-order digital filters in the presence of quantization and adder overflow. We consider filters employing saturation arithmetic; that is, we define the function f of Section II by

$$f(\xi) = \begin{cases} -1 & \text{for } \xi < -1 \\ \xi & \text{for } -1 \leqq \xi \leqq 1 \\ 1 & \text{for } \xi > 1. \end{cases} \tag{9}$$

The following theorem is fundamental to our analysis.

Theorem 1: Let the matrix A be defined by (1b) *in which the values of a, b are specified by some point lying within the open triangular region of Fig.* 2 *(determined by the straight lines: $b \pm a = 1$, $b = -1$). Let the mapping F be defined by* (3) *in which the function f is specified by* (9). *Then, for any sequence $\nu^{(k+1)}$, $k = 0, 1, 2, \cdots$, satisfying* (8) *for some ϵ in the interval $0 \leqq \epsilon < 1$ and some nonnegative integer K, the solution of* (7) *satisfies $\lim_{k\to\infty} \| z^{(k)} \| = 0$ provided that there exists a real number σ such that*

$$1 - \sigma^2 a^2 > 0, \tag{10}$$

$$[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2 > 0, \tag{11}$$

and

$$\sqrt{1 - \sigma^2 a^2} + \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2} > |b^2 - (1 - \sigma)a^2|. \tag{12}$$

The proof of Theorem 1 is given in the Appendix. We now seek to determine those points lying within the triangular region of Fig. 2 which specify values of a, b such that the inequalities (10), (11), (12) are satisfied for some value of σ.

We begin by examining the case in which $\sigma = 0$. In this case (10) is satisfied for all a, b and, as shown in the Appendix, (11) is satisfied for only those values of a, b specified by points lying within the open crosshatched region of Fig. 5 which, for $\sigma = 0$, is the open crosshatched region of Fig. 3. By squaring each side of the inequality (12), it is easily shown that that inequality, with $\sigma = 0$, is equivalent to

$$(1 - a^2 - b^2) + \sqrt{(1 - a^2 - b^2)^2 - 4a^2b^2} > 0.$$

Since values of a, b specified by points lying within the crosshatched region of Fig. 3 satisfy $1 - a^2 - b^2 > 0$, it is clear that all such values (and only those values) of a, b satisfy (10), (11), and (12) for $\sigma = 0$.

For negative values of σ and for $\sigma \geqq 2$ it is clear that the crosshatched region of Fig. 5 lies interior to the crosshatched region of Fig. 3. Thus, consideration of such values of σ can determine no values of a, b that are not already determined in Fig. 3 by consideration of the $\sigma = 0$ case.

We now show that values of σ in the interval $1 \leqq \sigma < 2$ yield no values of a, b satisfying (10), (11), and (12) that cannot also be determined by considering some value of σ in the interval $0 < \sigma \leqq 1$. Let $\hat{\sigma}$ satisfy $1 \leqq \hat{\sigma} < 2$ and then define $\tilde{\sigma} = 2 - \hat{\sigma}$. Clearly $0 < \tilde{\sigma} \leqq 1$. Now, if (10) is satisfied for $\sigma = \hat{\sigma}$, then, clearly, (10) is also satisfied for $\sigma = \tilde{\sigma}$. The expression on the left-hand side of (11) can be rewritten as $(1 - b^2)^2 + a^2\{[a^2 - 2(1 + b^2)] - [a^2 - (1 - b)^2]\sigma(2 - \sigma)\}$. The form of this expression shows that it has the same value for $\sigma = \hat{\sigma}$ and $\sigma = \tilde{\sigma}$. Finally, it is clear that if (12) is satisfied for $\sigma = \hat{\sigma}$, then (12) is also satisfied for $\sigma = \tilde{\sigma}$ since [using our observations regarding (10) and (11)] the left-hand side of that inequality is not decreased by

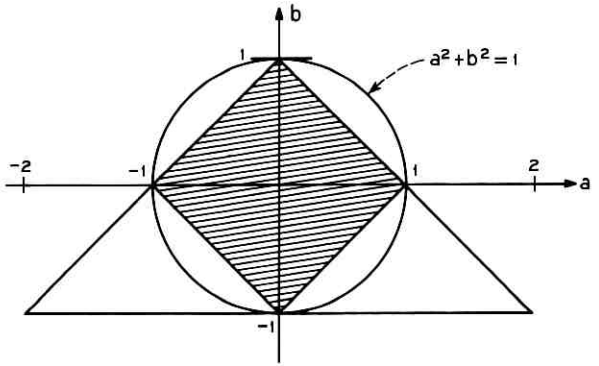

Fig. 3—Region in which inequality (11) is satisfied for $\sigma = 0$.

replacing $\hat{\sigma}$ with $\bar{\sigma}$, and since

$$| b^2 - (1 - \hat{\sigma})a^2 | = | b^2 + (1 - \bar{\sigma})a^2 | \geqq | b^2 - (1 - \bar{\sigma})a^2 |.$$

There remains to consider only those values of σ in the interval $0 < \sigma \leqq 1$. Thus, for each such value of σ we wish to determine the values of the parameters a, b specified by points lying within the open crosshatched region of Fig. 5 and, from (10), within the open region specified by $| a | < 1/\sigma$, for which the inequality (12) is satisfied. It is not difficult to show that for each value of σ satisfying $0 < \sigma \leqq 1$ the function

$$\begin{aligned}\varphi_\sigma(a, b) &\equiv \sqrt{1 - \sigma^2 a^2} + \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2} \\ &\quad - |b^2 - (1 - \sigma)a^2|,\end{aligned}$$

whose domain is that portion of the open crosshatched region of Fig. 5 where $| a | < 1/\sigma$, is monotone decreasing in $| a |$ for each value of b in the interval $-1 < b \leqq 0$. Thus, the region in which the inequality (12) is satisfied is easily located by determining the curves $\varphi_\sigma(a, b) = 0$. Moreover, because of the above observation concerning the monotonicity of φ_σ, it is easy to determine these curves numerically. Several such curves, for various values of σ, are shown in Fig. 4.

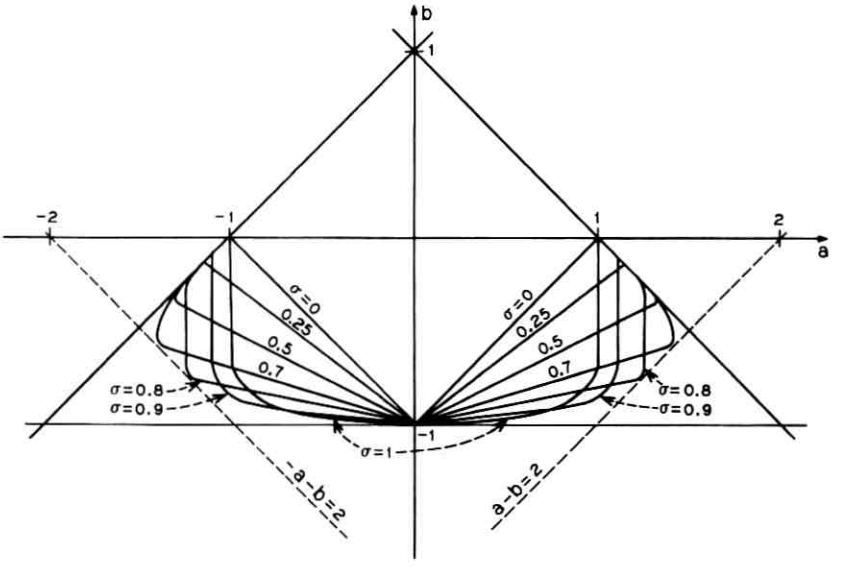

Fig. 4—Location of the $\varphi_\sigma(a, b) = 0$ curves for several values of σ.

The region in which the inequalities (10), (11), (12) are satisfied for *some* real number, σ, is the union of the regions determined by the $\varphi_\sigma(a, b) = 0$ curves for $0 \leqq \sigma \leqq 1$. The numerical results show that this region has the shape indicated by the crosshatched area in Fig. 2. The boundary of this region appears to be determined by the straight lines $b \pm a = 1$ for $b \geqq -\frac{1}{3}$, and by the ellipse $a^2 + 8b(1 + b) = 0$ for $b \leqq -\frac{1}{3}$.

It is clear that there are several ways in which our analysis could be refined in order to provide the possibility of improving upon the result of Theorem 1. In the next section, however, we show a fundamental limitation on the extent of any such improvement. We show there, how to construct counterexamples which demonstrate that for a certain large portion of the uncrosshatched area of the triangular region of Fig. 2 (in particular, the shaded areas in each lower corner) the conclusion of Theorem 1 is, in fact, false.

IV. COUNTEREXAMPLES

We now show how to construct the counterexamples which have been referred to in the preceding sections. We begin by showing that, when the function f is defined by (9), and when the values of the filter's coefficients are determined by any point lying within the open shaded regions in each lower corner of the triangle shown in Fig. 2, then there exist nonzero initial conditions and, for some $\epsilon < 1$, a periodic input sequence $v^{(k+1)}$ satisfying (8) such that the solution of (7) is periodic (and thus does not satisfy $\lim_{k\to\infty} \| z^{(k)} \| = 0$).

We first consider values of a, b determined by any point lying within the shaded open region in the lower *left* corner of the triangle of Fig. 2. In particular, we assume that

$$ab > 1. \tag{13}$$

It is also clear that the inequality

$$a < -b^2 \tag{14}$$

holds for any such point. Denoting the initial condition $z^{(0)}$ by $z^{(0)} = (x^{(0)}, y^{(0)})^T = (p, q)^T$, and considering an input sequence having period *three*, specified by $v^{(1)} = 0$, $v^{(2)} = 1 - p$, $v^{(3)} = -1 - q$, it is clear from Table I that (7) has a nonzero periodic solution provided that values of p, q can be found such that the inequalities specified in parentheses in Table I are satisfied.

The inequalities in the $v^{(k+1)}$ column [which must hold in order that

the input sequence satisfy (8) for some $\epsilon < 1]$ and the inequalities on the first line of the column labeled "$bx^{(k)} + ay^{(k)} + v^{(k+1)}$" are equivalent to:

$$0 < p < 2,$$
$$-2 < q < 0,$$
$$-1 < bp + aq < 1. \tag{15}$$

Thus, so long as we consider only positive values of p and negative values of q, these inequalities will always hold whenever p and q have sufficiently small magnitude. The remaining two inequalities specified in Table I will be satisfied provided that

$$(1 - ab)p - (a^2 + b)q < 0$$

and

$$(a + b^2)p - (1 - ab)q < 0. \tag{16}$$

In view of the inequalities (13), (14), it is clear that there exist values of $p > 0$ and $q < 0$ such that (16) is satisfied. Moreover, it is clear that the magnitudes of p and q can be scaled such that the inequalities (15) are also satisfied. Thus, there exists a nonzero periodic (of period three) solution of (7).

For values of a, b determined by any point lying within the open shaded region in the lower *right* corner of the triangle of Fig. 2 a similar line of reasoning shows that a nonzero solution of period *six* can be obtained. The existence of such a solution is easy to demonstrate by noting the odd-symmetry of the function f and, with the initial condition $z^{(0)} = (p, q)^T$, showing that, with $v^{(1)} = 0$, $v^{(2)} = 1 + p$, $v^{(3)} = -1 + q$, there exist values of p, q such that $z^{(3)} = (-p, -q)^T$. We omit the details.

The above procedure for constructing counterexamples is concerned explicitly with the solutions of (7). The simple relationships between (7) and the original equations of interest, i.e., eqs. (1) and (2), are described

TABLE I—CONSTRUCTION OF A PERIODIC SOLUTION FOR EQUATION (7)

k	$x^{(k)}$	$y^{(k)}$	$v^{(k+1)}$	$bx^{(k)} + ay^{(k)} + v^{(k+1)}$	$f - v^{(k+1)}$
0	p	q	0	$(-1<)\ bp + aq\ (<1)$	$bp + aq$
1	q	$bp + aq$	$(-1<)\ 1 - p\ (<1)$	$bq + abp + a^2q + 1 - p\ (>1)$	p
2	$bp + aq$	p	$(-1<)\ -1 - q\ (<1)$	$b^2p + abq + ap - 1 - q\ (<-1)$	q
3	p	q	0	...	...

in Section II. It is instructive, however, to consider explicitly the implications of such counterexamples concerning the solutions of (1) and (2).

Let any values of the parameters a, b, determined by some point lying within the open shaded regions in the lower corners of the triangle of Fig. 2, be given. Consider *any* counterexample constructed according to the above procedure. Then, assuming that the quantization occurs in an appropriate manner (or, assuming that there is no quantization) it is a straightforward matter to use the relationships between the variables of (1), (2), and (7) to demonstrate a periodic input sequence $u^{(k)}$ and appropriate values of the initial conditions $v^{(-1)}$, $v^{(0)}$ such that the response of the linear model of the filter of Fig. 1 [i.e., $w^{(k)}$, the solution of (1)] is asymptotic to a periodic sequence, and satisfies $\| w^{(k)} \| < 1$ for all sufficiently large k, while the response of the nonlinear filter [i.e., $r^{(k)}$, the solution of (2)], although also periodic, is such that $\| w^{(k)} - r^{(k)} \|$ does not approach zero as $k \to \infty$.

These counterexamples, while clearly demonstrating that there exists *potential* trouble whenever a filter's coefficients are assigned such "bad" values, do not show that such behavior will necessarily be possible for some *particular* filter. They do not demonstrate, for example, that with a particular (specified) kind of quantization, and with a particular set of permissible values for the filter's input sequence, there will necessarily exist a periodic input sequence for which the linear, and the nonlinear digital filters have asymptotically different responses. It is possible, however, by considering at the outset the details of the quantization and thereby imposing somewhat different constraints (to those of Table I) on the values chosen for p, q, $\nu^{(1)}$, to construct certain counterexamples which show just that.

We assume that the values specified for the parameters a, b are determined by a point lying within the open shaded region in the lower *left* corner of the triangle of Fig. 2. (A similar development could, of course, be considered for the other shaded region.) We also assume that a certain *finite* set Q of allowable *machine numbers*, satisfying $x \in Q \Rightarrow | x | \leqq 1$, is specified. Thus, we assume that for the nonlinear digital filter with quantization the variables $u^{(k)}$, $v^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ of Fig. 1 can assume only those values specified by the set Q. Furthermore, we assume that the filter employs saturation arithmetic with the overflow and quantization effects both specified by a certain function f_q; that is, given any values for $u^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ taken from the set Q, the value for $v^{(k)}$ appearing at the output of the summation point in Fig. 1 is specified by

$$v^{(k)} = f_q(av^{(k-1)} + bv^{(k-2)} + u^{(k)}). \tag{17}$$

If, for example, with $a = -1.3$ and $b = -0.9$, the values $u^{(k)} = 0.0$, $v^{(k-1)} = -0.9$, $v^{(k-2)} = 1.0$ are considered; and if the quantization is accomplished by simply rounding the ideal output of the summation point to the nearest tenth, the value $v^{(k)}$ specified by (17) is $v^{(k)} = 0.3$.

Clearly, if the set Q imposes sufficiently severe (indeed, for practical purposes, *unreasonable*) restrictions on the values that the input sequence and the initial conditions may assume, then it will be impossible to construct a counterexample. It is no surprise, therefore, that the success of the process to be described depends upon the assumption that the quantization is "sufficiently fine" (that is, that there are sufficient quantization levels distributed throughout the interval $[-1, 1]$), and that when $| av^{(k-1)} + bv^{(k-2)} + u^{(k)} | \leqq 1$, the actual output of the summation point is reasonably close to the ideal value, that is,

$$f_q(av^{(k-1)} + bv^{(k-2)} + u^{(k)}) \approx av^{(k-1)} + bv^{(k-2)} + u^{(k)}. \tag{18}$$

We first note that the values of a, b determined by any point lying within the open shaded region in the lower left corner of the triangle of Fig. 2 are such that $-1 < a - b < 1$. Thus, since $b/a > 0$, we also have $-1 < a - b + b/a$ and $a - b < 1$. This ensures that the open intervals $(-1, 1)$ and $(a - b, a - b + b/a)$ overlap. Hence, if the quantization is sufficiently fine, there exists $u^{(1)} \in Q$ such that

$$-a/b < 0 < b - a + u^{(1)} < b/a < 1. \tag{19}$$

Thus, for such a value of $u^{(1)}$,

$$b - a(b - a + u^{(1)}) < 0$$

and

$$a + b(b - a + u^{(1)}) < 0.$$

Hence, for sufficiently fine quantization, there exist $u^{(2)}, u^{(3)} \in Q$ such that

$$1 + b - a(b - a + u^{(1)}) - u^{(2)} < 0, \tag{20}$$

and

$$1 + a + b(b - a + u^{(1)}) + u^{(3)} < 0. \tag{21}$$

We let $r_{\max}$ denote the most positive value in the set Q and let $r_{\min}$ denote the most negative value in Q. We also let

$$r_2^{(1)} = f_q(ar_{\min} + br_{\max} + u^{(1)}),$$

$$r_2^{(2)} = r_{\max},$$

$$r_2^{(3)} = r_{\min}.$$

Now, assuming that (18) holds, it can be expected, due to (20) and (21), that there exist p, q such that

$$(1 - ab)p - (a^2 + b)q = 1 - br_{\min} - ar_2^{(1)} - u^{(2)} < 0, \quad (22)$$

$$(a + b^2)p - (1 - ab)q = 1 + ar_{\max} + br_2^{(1)} + u^{(3)} < 0. \quad (23)$$

Moreover, if the quantization is sufficiently fine, the values for $u^{(2)}$ and $u^{(3)}$ can be chosen such that $1 - br_{\min} - ar_2^{(1)} - u^{(2)} \approx 0$ and $1 + ar_{\max} + br_2^{(1)} + u^{(3)} \approx 0$, and such that the values of these expressions are in the proper ratio that, in fact, *small* values of $p > 0$ and $q < 0$ are determined by the equations in (22), (23). Thus, since for sufficiently fine quantization

$$ar_{\min} + br_{\max} + u^{(1)} \approx b - a + u^{(1)}, \quad (24)$$

and, due to (19), it is reasonable to expect that there exists $\nu^{(1)}$ such that

$$-1 < bp + aq + \nu^{(1)} = ar_{\min} + br_{\max} + u^{(1)} < 1. \quad (25)$$

Furthermore, for $p > 0$, $q < 0$ small, we expect that the following inequalities also hold:

$$-1 < r_2^{(1)} - bp - aq < 1, \quad (26)$$

$$-1 < r_{\max} - p < 1, \quad (27)$$

$$-1 < r_{\min} - q < 1. \quad (28)$$

Assuming therefore that the values of $u^{(1)}$, $u^{(2)}$, $u^{(3)}$, p, q, $\nu^{(1)}$ are such that (22), (23), (25), (26), (27), and (28) hold, we proceed with the construction of a counterexample by simply assigning the values to the remaining variables that are dictated by the relationships specified in Section II. In particular, we let

$$s_2^{(1)} = r_2^{(1)} - bp - aq, \qquad e^{(1)} = s_2^{(1)} - \nu^{(1)},$$

$$s_2^{(2)} = r_{\max} - p, \qquad e^{(2)} = -1 + r_{\max},$$

$$s_2^{(3)} = r_{\min} - q, \qquad e^{(3)} = 1 + r_{\min},$$

$$\nu^{(2)} = 1 - p,$$

$$\nu^{(3)} = -1 - q,$$

and

$$r_1^{(1)} = r_2^{(3)}, \qquad s_1^{(1)} = s_2^{(3)},$$

$$r_1^{(2)} = r_2^{(1)}, \qquad s_1^{(2)} = s_2^{(1)},$$

$$r_1^{(3)} = r_2^{(2)}, \qquad s_1^{(3)} = s_2^{(2)}.$$

At this point, one final step remains in our construction of a counterexample. We have obtained periodic solutions of eqs. (4) and (2), with the solution of (4) satisfying $\| s^{(k)} \| < 1$. We would like to obtain the corresponding periodic sequence to which the solution of (1) is asymptotic. This sequence, which we shall call $\hat{w}^{(k)}$, is easily determined by the equations

$$\begin{bmatrix} \hat{w}_2^{(1)} \\ \hat{w}_2^{(2)} \\ \hat{w}_2^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & -b & -a \\ -a & 1 & -b \\ -b & -a & 1 \end{bmatrix}^{-1} \begin{bmatrix} u^{(1)} \\ u^{(2)} \\ u^{(3)} \end{bmatrix},$$

$$\hat{w}_1^{(1)} = \hat{w}_2^{(3)},$$

$$\hat{w}_1^{(2)} = \hat{w}_2^{(1)},$$

$$\hat{w}_1^{(3)} = \hat{w}_2^{(2)}.$$

We have now found a true counterexample only if the values of $\hat{w}^{(k)}$ are also such that $\| \hat{w}^{(k)} \| < 1$. It is reasonable to expect that this inequality will hold, however, since $\| \hat{w}^{(k)} - s^{(k)} \|$ is known to be small provided that the values of $e^{(1)}$, $e^{(2)}$, $e^{(3)}$ are small, and these terms will be small whenever the quantization is sufficiently fine [note that $e^{(1)} = r_2^{(1)} - bp - aq - \nu^{(1)}$, and recall the equality expressed in (25)].

Computer programs have been written which use the above process for constructing counterexamples, and which simulate the behavior of linear and nonlinear digital filters. It has been our experience, based upon experimentation with these programs, that counterexamples of the type described above can easily be found for values of the coefficients a, b determined by points lying within the shaded region in the lower left-hand corner of the triangle in Fig. 2 even when the quantization is extremely coarse, much coarser than the quantization occurring in current practical digital filter realizations. We give, for example, the following numerical counterexample, constructed according to the above procedure, in which we have intentionally considered very coarse quantization, and have also made the task even more difficult by choosing $u^{(1)}$, $u^{(2)}$, $u^{(3)}$ in, obviously, a somewhat less-than-optimum manner, with the result that $| p |$ and $| q |$ are larger than necessary. This does, however, cause the resulting sequences $r^{(k)}$, $w^{(k)}$ to be quite different.

We assume that the coefficient values $a = -1.3$, $b = -0.9$ have been specified. We also assume that

$$Q = \{-0.9, -0.8, \cdots, 0.9, 1\}.$$

We assume that the quantization is performed by simple rounding, at the output of the summation point, of the ideal sum to the nearest tenth. We then have $r_{\max} = 1$, $r_{\min} = -0.9$, and therefore, choosing

$$u^{(1)} = 0, \qquad u^{(2)} = 0.6, \qquad u^{(3)} = 0.2,$$

it follows that

$$r_2^{(1)} = 0.3, \qquad r_2^{(2)} = 1, \qquad r_2^{(3)} = -0.9. \tag{29}$$

We find that the approximate values of p, q, specified by (22), (23) are: $p = 0.711$, $q = -0.128$. Following the above outlined procedure, we find that all of the required relationships hold. The resulting periodic sequence $\hat{w}^{(k)}$ to which the sequence $w^{(k)}$ is asymptotic is specified by the following approximate values:

$$\hat{w}_2^{(1)} = 0.905, \qquad \hat{w}_2^{(2)} = 0.135, \qquad \hat{w}_2^{(3)} = -0.790. \tag{30}$$

Note that quite different solutions are specified by (29) and (30).

V. THE FORCED RESPONSE AND INPUT SCALING

We have shown in Section IV that the forced response of a stable second-order digital filter employing saturation arithmetic might not, for some inputs, be even asymptotically close to the filter's ideal response (the response of the linear filter) if the coefficients a, b are specified by a point lying outside the crosshatched region of the triangle in Fig. 2. More precisely, we have shown that this certainly happens for coefficient values determined by points lying within the shaded regions in each lower corner of that triangle (so long as certain reasonable assumptions hold concerning the nature of the quantization). Thus one concludes that, when designing a filter, it is desirable to avoid choosing such coefficient values. In practical applications, however, it might be the case that due to other considerations such a choice cannot be avoided. Then it is clear that the designer must be careful to impose appropriate restrictions on the filter's input sequence and on its initial conditions. He might, for example, scale the input sequence such that it is always small enough. The question thus arises: How small is "small enough"? One obvious answer to this question is that the input and the initial conditions be required to be small enough that the response of the *linear* filter [described by (1)] satisfies, for some $\delta > 0$ and all $k = 0, 1, 2, \cdots$, the inequality $\| w^{(k)} \| \leqq 1 - \delta$. Then, by using sufficiently many bits in the representation of the data, the quantization

error can always be made sufficiently small that the adder overflow nonlinearity is not encountered.

The results contained in a paper[3] on limit cycles can provide another answer to this question. This answer requires consideration of only the *asymptotic* nature of the input sequence, and applies to filters using a variety of kinds of arithmetic including, in particular, saturation arithmetic. It is clear from the analysis presented in Ref. 3, that it is sufficient that the input sequence $u^{(k+1)}$ and the quantization error sequence $e^{(k+1)}$ be such that the solution of (4) satisfy, for some nonnegative integer K, the inequality

$$|| s^{(k)} || + | e^{(k+1)} | < \delta, \qquad \text{for } k \geqq K,$$

where δ is one of the bounds specified in Theorem 1 of Ref. 3 for the sequence $\nu^{(k+1)}$. In the case of saturation arithmetic we have

$$\delta = \max \left\{ \frac{2 - |a|}{2 + |a|}, \frac{1 - |b|}{1 + |b|} \right\}.$$

Then, it is clear (by Theorem 1 of Ref. 3) that the solution of (7) satisfies $\lim_{k \to \infty} || z^{(k)} || = 0$, which therefore ensures proper asymptotic behavior of the forced response of the nonlinear filter.

VI. ACKNOWLEDGMENT

The author is pleased to acknowledge the helpful comments of his colleagues J. F. Kaiser and I. W. Sandberg concerning this work.

APPENDIX

The proof of Theorem 1 follows. The proof uses the following well-known result concerning the application of Liapunov's "second method" to the study of the stability of difference equations.[5-7]

Lemma 1: Let G denote a subset of the n-dimensional Euclidean space E^n containing the origin θ. If there exist continuous functions $W: G \to E^1$, $V: G \to E^1$, and if there exists a nonnegative integer K such that:

(i) $W(z) > 0$ *for all* $z \in G$, $z \neq \theta$,
(ii) $W(\theta) = 0$,
(iii) $V(z) \geqq 0$ *for all* $z \in G$,
(iv) $\Delta V(k, z) = V(g(k, z)) - V(z) \leqq -W(z)$ *for all* $k \geqq K$ *and all* $z \in G$,

then each solution of the difference equation $z^{(k+1)} = g(k, z^{(k)})$ which remains in G for all $k \geqq K$ approaches the origin as $k \to \infty$.

For any particular application, the effectiveness of Liapunov's method is of course highly dependent upon the appropriateness of the particular Liapunov function V that is chosen. The quadratic form that will now be described is quite useful for our purposes.

For any given values of the parameters a, b which specify an asymptotically stable linear digital filter, and with the eigenvalues of the matrix A of (1b) being denoted by λ_1, λ_2, let the Liapunov function V be defined by

$$V(z) = z^T B z, \tag{31}$$

with

$$B = \begin{bmatrix} |\lambda_1|^2 + |\lambda_2|^2 + 2\mu & -\sigma a \\ -\sigma a & 2 \end{bmatrix}, \tag{32}$$

where the values of σ and μ are yet to be determined.

In the following lemma we determine, for any given value of σ, those values of μ for which the matrices B and $B - A^T B A$ are positive definite.

Lemma 2: Let σ be a given real number. Then, necessary and sufficient conditions for the matrices B and $B - A^T B A$ both to be positive definite for values of a, b which specify an asymptotically stable linear digital filter are: that the values of the parameters a, b be restricted to those values specified by points lying within the open crosshatched region of Fig. 5, and that, with $\mu_1 < \mu_2$ specified by

$$\mu_{1,2} = \tfrac{1}{2}\{1 + b^2 - (1 - \sigma)a^2 - (|\lambda_1|^2 + |\lambda_2|^2) \pm \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2}\}, \tag{33}$$

a value be assigned to μ such that

$$\mu_1 < \mu < \mu_2 .$$

Proof: It is clear that a necessary and sufficient condition for the matrix B of (32) to be positive definite is that $\det B > 0$, which is equivalent to the inequality

$$\mu > -\tfrac{1}{2}(|\lambda_1|^2 + |\lambda_2|^2) + \tfrac{1}{4}\sigma^2 a^2. \tag{34}$$

The matrix $B - A^T B A$, which has the form

$$\begin{bmatrix} (|\lambda_1|^2 + |\lambda_2|^2) - 2b^2 + 2\mu & -a[\sigma + (2 - \sigma)b] \\ -a[\sigma + (2 - \sigma)b] & 2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2(1 - \sigma)a^2 - 2\mu \end{bmatrix},$$

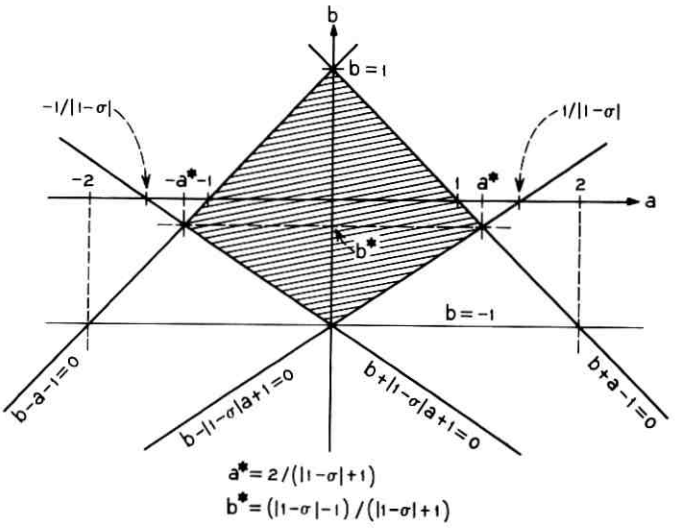

Fig. 5—Region of the a-b plane in which the matrices B and $B - A^TBA$ both may be positive definite.

is positive definite if and only if det $(B - A^TBA) > 0$ and

$$\mu > b^2 - \tfrac{1}{2}(|\lambda_1|^2 + |\lambda_2|^2). \tag{35}$$

As is easily verified, the inequality det $(B - A^TBA) > 0$ is equivalent to

$$\begin{aligned} &-4\mu^2 + 4[1 + b^2 - (1 - \sigma)a^2 - (|\lambda_1|^2 + |\lambda_2|^2)]\mu \\ &+ \{-(|\lambda_1|^2 + |\lambda_2|^2)^2 + 2[1 + b^2 - (1 - \sigma)a^2](|\lambda_1|^2 + |\lambda_2|^2) \\ &- a^2[\sigma + (2 - \sigma)b]^2 - 4b^2[1 - a^2(1 - \sigma)]\} > 0. \end{aligned} \tag{36}$$

We view the left-hand side of the inequality (36) as a quadratic function in the variable μ whose coefficients depend upon the values of the parameters σ, a, b. Clearly, for any choice of these parameter values, (36) will not be satisfied for all large values of $|\mu|$. Thus, a necessary and sufficient condition for the existence of real values of μ satisfying (36) is that the quadratic function on the left-hand side of (36) have distinct real zeros $\mu_1 < \mu_2$. Moreover, if such is the case, (36) will be satisfied if and only if $\mu_1 < \mu < \mu_2$. The zeros μ_1, μ_2 are given by (33) and, therefore, they are real and distinct if and only if

$$|1 - b^2 - (1 - \sigma)a^2| > |a| \cdot |\sigma + (2 - \sigma)b|. \tag{37}$$

We now prove that for any given value of σ, the values of the param-

eters a, b specified by points lying within the open triangular region of Fig. 2, and which satisfy (37), are those (and only those) values of a, b specified by points lying within the open crosshatched region of Fig. 5.

We begin by first showing that there exist no such values of a, b for which $1 - b^2 - (1 - \sigma)a^2 < 0$. Let us assume that this inequality holds for some value of σ. Then, since $1 - b^2 > 0$, it follows that $\sigma < 1$. Now, either

$$-\sigma/(2 - \sigma) \leqq b < 1, \tag{38}$$

or else

$$-1 < b < -\sigma/(2 - \sigma). \tag{39}$$

If (38) holds, then (37) is equivalent to

$$-1 + b^2 + (1 - \sigma) \mid a \mid^2 > \mid a \mid [\sigma + (2 - \sigma)b],$$

or

$$(b - \mid a \mid - 1)[b - (1 - \sigma) \mid a \mid + 1] > 0. \tag{40}$$

If, however, (39) holds, then (37) is equivalent to

$$(b + \mid a \mid - 1)[b + (1 - \sigma) \mid a \mid + 1] > 0. \tag{41}$$

By considering first only nonnegative values of a, and then considering only nonpositive values of a, it is easy to use Fig. 5 and, by inspection, determine that there exist no values of the parameters a, b specified by points lying within the triangular region, such that both (38) and (40) hold. Similarly, it is easy to verify that the same is true regarding inequalities (39) and (41).

We now assume that the parameters σ, a, b are to be chosen such that $1 - b^2 - (1 - \sigma)a^2 \geqq 0$. Then there are three cases to consider:

If $\sigma \geqq 1$, it follows that $\sigma + (2 - \sigma)b > 0$ and hence (37) is easily shown to be equivalent to

$$(b + \mid a \mid - 1)[b - \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{42}$$

If $\sigma < 1$ and (38) holds, then it follows that (37) is equivalent to

$$(b + \mid a \mid - 1)[b + \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{43}$$

If $\sigma < 1$ and (39) holds, then it follows that (37) is equivalent to

$$(b - \mid a \mid - 1)[b - \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{44}$$

By first considering only nonnegative values of a, and then considering only nonpositive values of a, it is easy to use Fig. 5 and, by inspection, determine that the inequality (42) is satisfied if and only if the values of the parameters a, b are determined by points lying within the open crosshatched region of Fig. 5. Similarly, the inequalities (38) and (43), or the inequalities (39) and (44), hold if and only if the values of the parameters a, b are determined by points lying within the open crosshatched region of Fig. 5.

It can easily be shown that for any given value of σ, and any values of the parameters a, b specified by points lying within the open crosshatched region of Fig. 5, it follows from $\mu > \mu_1$ that the inequalities (34) and (35) also hold. We omit the details of the algebra. □

Proof of Theorem 1: Let the Liapunov function V be defined for all $z \in G \equiv E^2$ by (31) and (32) with the values of σ, a, b, μ assumed to be such that both of the matrices B and $B - A^TBA$ are positive definite. It is clear that the equations

$$z^T(B - A^TBA)z = c, \qquad c > 0 \tag{45}$$

define a family of concentric ellipses, centered at the origin θ in the x-y plane [where $z = (x, y)^T$]. The origin also lies between the two parallel straight lines $bx + ay = \pm(1 - \epsilon)$, each of which is tangent to exactly one (in fact, the same one) of the ellipses (45). Thus, there is a unique value of $c^* > 0$ such that

$$c^* = \min \{z^T(B - A^TBA)z : bx + ay = \pm(1 - \epsilon)\}.$$

Let the function W be defined for all $z \in G \equiv E^2$ by

$$W(z) = \min \{z^T(B - A^TBA)z, c^*\}.$$

Thus, $W(z)$ is defined by the positive definite quadratic form $z^T(B - A^TBA)z$ for all points lying within the ellipse $z^T(B - A^TBA)z = c^*$, and $W(z)$ is defined by $W(z) = c^*$ for all other points in the x-y plane.

It is clear that for each value of $\nu^{(k+1)}$ for which (8) holds, the points of the x-y plane determined by $bx + ay + \nu^{(k+1)} \geqq 1$ lie on the opposite side of the line $bx + ay = 1 - \epsilon$ from the ellipse $z^T(B - A^TBA)z = c^*$. The situation is similar regarding the points of the x-y plane determined by $bx + ay + \nu^{(k+1)} \leqq -1$ and the line $bx + ay = -(1 - \epsilon)$. See Fig. 6.

With

$$g(k, z) \equiv F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix},$$

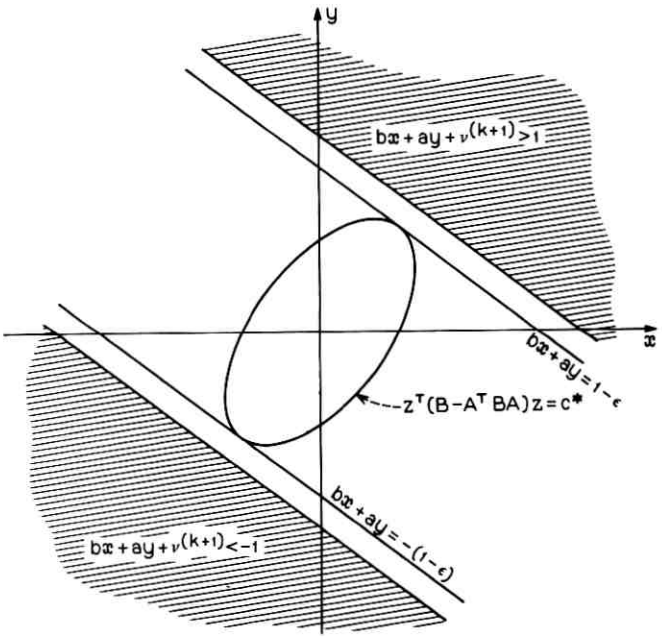

Fig. 6—Location of the ellipse $z^T(B - A^TBA)z = c^*$.

it follows that

$$\Delta V(k, z) = V(g(k, z)) - V(z)$$

$$= \left[F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right]^T B\left[F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right] - z^TBz.$$

Thus, whenever $| bx + ay + \nu^{(k+1)} | \leqq 1$, we have

$$\Delta V(k, z) = -z^T(B - A^TBA)z \leqq -W(z).$$

When $bx + ay + \nu^{(k+1)} > 1$,

$$\Delta V(k, z) =$$
$$- \{[(|\lambda_1|^2 + |\lambda_2|^2) + 2\mu]x^2 - 2\sigma axy + [2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2\mu]y^2$$
$$+ 2\sigma a(1 - \nu^{(k+1)})y - 2(1 - \nu^{(k+1)})^2\}; \tag{46}$$

and when $bx + ay + \nu^{(k+1)} < -1$,

$$\Delta V(k, z) =$$
$$- \{[(|\lambda_1|^2 + |\lambda_2|^2) + 2\mu]x^2 - 2\sigma axy + [2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2\mu]y^2$$
$$- 2\sigma a(1 + \nu^{(k+1)})y - 2(1 + \nu^{(k+1)})^2\}. \tag{47}$$

It is an elementary result of analytic geometry that a general second-degree equation of the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$ represents an ellipse if and only if $b^2 - 4ac < 0$. It follows that, if we consider the constant-ΔV loci in the $bx + ay + \nu^{(k+1)} > 1$ region, and in the $bx + ay + \nu^{(k+1)} < -1$ region of the x-y plane, a necessary and sufficient condition for these loci to be arcs of concentric ellipses is:

$$4\mu^2 - 4[1 - (|\lambda_1|^2 + |\lambda_2|^2)]\mu + [\sigma^2 a^2 - 2(|\lambda_1|^2 + |\lambda_2|^2) + (|\lambda_1|^2 + |\lambda_2|^2)^2] < 0. \quad (48)$$

Furthermore, since $\Delta V(k, z)$ is continuous in z, and since the values of $\Delta V(k, z)$ along the lines $bx + ay + \nu^{(k+1)} = \pm 1$ are given by $\Delta V(k, z) = -z^T(B - A^TBA)z$, with $B - A^TBA$ a positive definite matrix, it is clear that when (48) is satisfied, the constant-ΔV curves specified by (46) (temporarily extending the domain of definition of that function to the entire x-y plane) are of the type shown in either Fig. 7a or Fig. 7b; that is, the line $bx + ay + \nu^{(k+1)} = 1$ intersects only certain constant-ΔV curves—in particular, only certain such curves for which the value of ΔV is negative. Thus, the center of the ellipses is situated to one side or the other of the line $bx + ay + \nu^{(k+1)} = 1$ in such a manner that the constant-ΔV ellipses for which ΔV is positive are not intersected by the line. Considering, however, that when $\Delta V(k, z)$ of (46) is evaluated at $z = \theta$ its value is positive, it is clear that Fig. 7b is impossible. Thus [applying exactly the same reasoning to the constant-ΔV curves defined by (47)], it follows that whenever the inequality (48) is satisfied, the

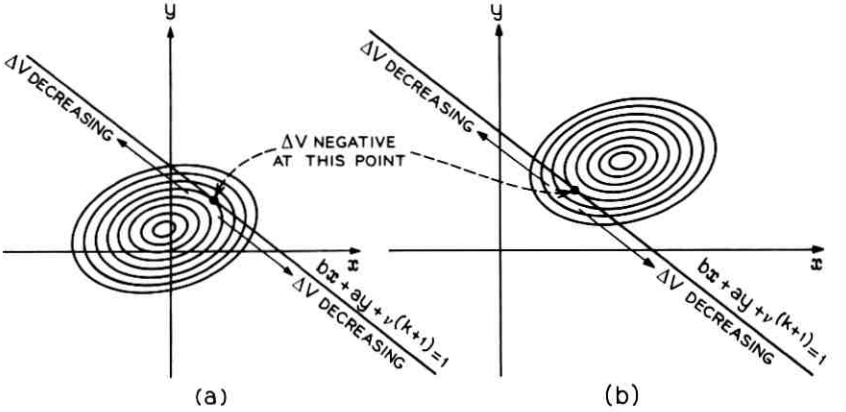

Fig. 7—Possible shape of constant-ΔV curves defined by equation (46).

function $\Delta V(k, z)$ achieves its maximum for $bx + ay + \nu^{(k+1)} \geqq 1$ on the line $bx + ay + \nu^{(k+1)} = 1$, and similarly for the behavior of $\Delta V(k, z)$ in the $bx + ay + \nu^{(k+1)} \leqq -1$ region of the x-y plane. It follows, therefore, that there exists $c' \geqq c^* > 0$ such that for $bx + ay + \nu^{(k+1)} \geqq 1$,

$$\Delta V(k, z) \leqq -c' \leqq -c^* = -W(z). \tag{49}$$

Similarly, there exists $c'' \geqq c^* > 0$ such that for $bx + ay + \nu^{(k+1)} \leqq -1$,

$$\Delta V(k, z) \leqq -c'' \leqq -c^* = -W(z). \tag{50}$$

We have shown that, with the functions V, W defined as specified above, the hypotheses of Lemma 1 are satisfied. Thus, the solution of (7) satisfies $\lim_{k\to\infty} \| z^{(k)} \| = 0$ provided that the values of σ, a, b, μ are such that B and $B - A^T BA$ are positive definite, and provided that (48) holds.

We view the left-hand side of the inequality (48) as a quadratic function in the variable μ whose coefficient values depend upon the values of the parameters σ, a, b. Clearly, for any choice of these parameter values (48) will not hold for all large values of $|\mu|$. Thus, a necessary and sufficient condition for the existence of real values of μ satisfying (48) is that the quadratic function on the left-side of (48) has distinct real zeros $\hat{\mu}_1 < \hat{\mu}_2$; moreover (48) will be satisfied if and only if $\hat{\mu}_1 < \mu < \hat{\mu}_2$. The zeros $\hat{\mu}_1$, $\hat{\mu}_2$ are given by

$$\hat{\mu}_{1,2} = \tfrac{1}{2}[1 - (|\lambda_1|^2 + |\lambda_2|^2) \pm \sqrt{1 - \sigma^2 a^2}]. \tag{51}$$

They are real and distinct if and only if the inequality (10) holds.

According to Lemma 2, for any given value of σ the matrices B and $B - A^T BA$ are positive definite for values of a, b that are specified by some point lying within the open triangular region of Fig. 2 if and only if $\mu_1 < \mu < \mu_2$, where μ_1, μ_2 are specified by (33). Thus, assuming that σ, a, b satisfy (10) and (11), there exists a value of μ such that B and $B - A^T BA$ are positive definite and such that (48) holds if and only if the open intervals (μ_1, μ_2) and $(\hat{\mu}_1, \hat{\mu}_2)$ overlap. That is, if and only if $\mu_1 < \hat{\mu}_2$ and $\hat{\mu}_1 < \mu_2$. Using (33) and (51), these last two inequalities are easily shown to be equivalent to (12). □

REFERENCES

1. Sandberg, I. W., "A Theorem Concerning Limit Cycles in Digital Filters," Proc. Seventh Annual Allerton Conf. on Circuit and System Theory, (October 1969), pp. 63–68.
2. Ebert, P. M., Mazo, J. E., and Taylor, M. G., "Overflow Oscillations in Digital Filters," B.S.T.J., *48*, No. 9 (November 1969), pp. 2999–3020.

3. Willson, A. N., Jr., "Limit Cycles due to Adder Overflow in Digital Filters," to be published in IEEE Trans. on Circuit Theory, *CT-19*, No. 4 (July 1972).
4. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. on Audio and Electroacoustics, *AU-16*, No. 3 (September 1968), pp. 413–421.
5. Hahn, W., *Theory and Application of Liapunov's Direct Method*, Englewood Cliffs, N. J.: Prentice-Hall, 1963, pp. 146ff.
6. Freeman, H., *Discrete-Time Systems*, New York: Wiley, 1965, pp. 158ff.
7. Hurt, J., "Some Stability Theorems for Ordinary Difference Equations," SIAM J. Numer. Anal., *4*, No. 4 (December 1967), pp. 582–596.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Three-Dimensional Small-Signal Analysis of Bipolar Transistors

By J. L. BLUE

(Manuscript received November 15, 1971)

One-dimensional transistors are well-understood today; computer techniques for detailed large-signal and small-signal analyses are available. Real transistors are three-dimensional, however, and lateral effects are only understood qualitatively. Accurate modeling of lateral effects cannot be accomplished without quantitative analyses of three-dimensional transistors. Unfortunately, even the simplest analysis of lateral effects leads to a partial differential equation. In this paper, a fast and accurate numerical technique is used to solve the partial differential equation. This makes feasible a three-dimensional small-signal analysis of transistors operating in the low-injection regime.

Calculated h-parameters for a high-frequency, double-diffused, silicon transistor are in good agreement with experimental values.

I. INTRODUCTION

One-dimensional transistors are well-understood today; computer techniques for detailed large-signal and small-signal analyses are available.[1] A new charge-control model,[2] which is quite promising for use in circuit analysis programs, has been developed with the aid of insight obtained from the large-scale computer analyses.

However, real transistors are three-dimensional and lateral effects are only understood qualitatively. Accurate modeling of lateral effects cannot be accomplished without quantitative analyses of three-dimensional transistors. Unfortunately, even the simplest analysis of lateral effects leads to a partial differential equation [namely (7)]. Except for certain very simple transistor geometrical configurations, an analytic solution to the partial differential equation is not possible, and up to now, numerical solutions have been at best difficult and expensive in computer time.

In this paper, a fast and accurate numerical technique[3,4] is used to

solve the partial differential equation. This makes feasible a three-dimensional small-signal analysis of transistors operating in the low-injection regime.

It is hoped that with the aid of this technique, quantitative understanding of lateral effects in transistors can be achieved, and an accurate but simple model can be developed.

In Section II, the partial differential equation for the potential in the base region of the transistor is derived. Section III briefly reviews the solution technique. In Section IV, the calculation of h-parameters from the partial differential equation solution is described. Section V presents the results from a sample calculation, and compares them with experimental values.

II. TRANSISTOR MODEL

In this section, a discrete bipolar transistor with collector contact made to the substrate is analyzed. (A planar bipolar transistor could equally well have been chosen.) The small-signal behavior of the transistor in one dimension is modeled and a partial differential equation in the other two dimensions is derived.

Figure 1 is a cross section perpendicular to the surface of a typical discrete transistor with diffused base and emitter. A partial differential equation for φ, the potential in the base region, will now be derived. Let z be the coordinate perpendicular to the surface of the transistor. Let $\varphi(x, y)$ be the potential in the base, $\rho(x, y)$ be the charge density, and $\mathbf{J}(x, y)$ be the current density parallel to the transistor surface; φ, ρ, σ, and $\mathbf{J}$ are average values, and are taken to be independent of z.

Ohm's law is

$$\nabla\varphi = -\frac{1}{\sigma}\mathbf{J}, \tag{1}$$

where σ is the conductivity, and the continuity equation for the charge is

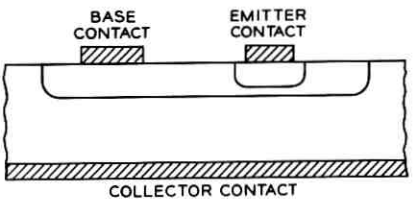

Fig. 1—Cross section of transistor.

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0. \tag{2}$$

The $\partial\rho/\partial t$ term is composed of capacitatively and resistively injected charge into the base, and will be considered shortly. Taking the divergence of (1) and substituting (2), one obtains

$$\nabla^2 \varphi = \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = \frac{1}{\sigma} \frac{\partial \rho}{\partial t}. \tag{3}$$

In order to calculate $\partial\rho/\partial t$, the net rate of the charge injection into the base region, a specific model is necessary. This paper will consider only a simple model, since the purpose is to illustrate the technique rather than to model exactly some particular transistor. Two separate sets of simplifying assumptions will be made, one set in calculating $1/\sigma\ \partial\rho/\partial t$ to set up the partial differential equation, and the second set in calculating small-signal parameters from the solution of the partial differential equation. The second set will be discussed in Section IV.

It is assumed that the emitter doping is sufficiently high that the potential on the emitter side of the emitter-base junction is constant. This constant is taken equal to zero. It is similarly assumed that the potential on the collector side of the collector-base junction is constant. The base region is divided into two parts—the *active base region*, directly under the emitter diffusion, and the *passive base region*, the remainder of the base. The conductivity of each of the two regions is taken to be constant, but the two conductivities are not equal. Finally, low injection is assumed so that current and charge injected by the emitter-base junction is uniform in the active base region.

With these assumptions, the rate of charge injection into the base may be calculated. Equations (4), (5), and (6) may be interpreted with the aid of the equivalent circuit of Fig. 2, which is essentially a hybrid-π model of a transistor neglecting base resistance corrections. The charge injected into the base at point (x, y) may be obtained by calculating the base current for the circuit in Fig. 2; such an equivalent circuit obtains at each point of the active base. Only the circuit elements which contribute to charge flow in the z-direction are included; the lateral charge flow is calculated by solving the partial differential equation (3). Thus, it is not necessary to include in the circuit of Fig. 2 the usual hybrid-π base resistances. In the equations below, $\varphi(x, y) = v_{BE}(x, y)$ is the only quantity that depends on x and y, for (x, y) in the active base region. The rate of charge injection into the base has three components.

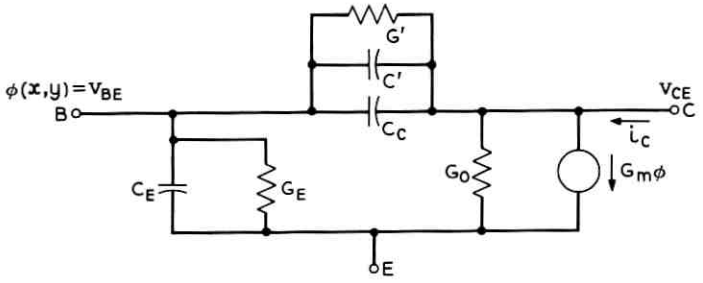

Fig. 2—Equivalent circuit used to calculate charge injection into the base and current injection into the collector, at a point (x, y) of the active base region.

(*i*) The dynamic resistance and capacitance of the emitter-base junction contributes to $\partial\rho/\partial t$ only in the active base region.

$$\left(\frac{\partial\rho}{\partial t}\right)_1 = \frac{1}{W_A}\left(C_E\frac{\partial\varphi}{\partial t} + G_E\varphi\right), \tag{4}$$

where W_A is the width (z-direction) of the active base region; C_E is the dynamic capacitance per unit area of the emitter-base junction, and includes the ordinary junction capacitance; and G_E is the conductance per unit area.

(*ii*) The collector junction capacitance contributes

$$\left(\frac{\partial\rho}{\partial t}\right)_2 = \frac{C_C}{W}\frac{\partial}{\partial t}(\varphi - v_{CE}), \tag{5}$$

where C_C is the capacitance per unit area of the collector-base junction, v_{CE} is the potential of the collector side of the collector-base junction, and W in the active base region is W_A and in the passive base region is W_P, the width of the passive base.

(*iii*) Base-width modulation in the active base region contributes

$$\left(\frac{\partial\rho}{\partial t}\right)_3 = \frac{1}{W_A}\left(G' + C'\frac{\partial}{\partial t}\right)(\varphi - v_{CE}), \tag{6}$$

where the conductance G' and capacitance C' are each per unit area. Generally C' and G' are not important. In addition, base-width modulation is the source of G_0, a conductance per unit area.

In the passive base region, the only charge injected is through the capacitance C_C.

Collecting the three terms, assuming frequency dependence $e^{i\omega t}$,

defining sheet resistance $R_A = 1/W_A\sigma_A$ and $R_P = 1/W_P\sigma_P$, the following partial differential equations for the potential are obtained.

Active base:

$$\nabla^2\varphi = R_A[G_E + G' + i\omega(C_C + C_E + C')]\varphi - R_A[G' + i\omega(C' + C_C)]v_{CE}\,. \tag{7}$$

Passive base:

$$\nabla^2\varphi = i\omega R_P C_C(\varphi - v_{CE}). \tag{8}$$

For typical parameters,

$$\begin{aligned} G_E &\gg G_0, \\ C_E &\gg C_C + C', \\ C_C &\gg C'; \end{aligned} \tag{9}$$

then in the active base the equation simplifies to

$$\nabla^2\varphi = R_A[(G_E + i\omega C_E)\varphi - (G' + i\omega C_C)v_{CE}]. \tag{10}$$

This equation was used for numerical calculations.

Equations (8) and (10) are of the form

$$\nabla^2\varphi = \gamma^2(\varphi - v_{CE}), \tag{11}$$

where γ^2 is frequency-dependent and takes on different values γ_A^2 and γ_P^2 in the active and passive base regions, but otherwise is independent of x and y.

Any other modeling of the z-direction of the transistor is equally usable, provided that it leads to equations of the type

$$\nabla^2\varphi = a\varphi + b, \tag{12}$$

where a and b are independent of x and y in each of several regions.

Figure 3 shows a typical double-base-stripe discrete bipolar transistor from the top. The active base region is under the emitter diffusion, marked E; this region is represented by eq. (10). Under the two regions marked B, φ is constant and $\varphi = v_{BE}$. The remainder of the region shown is the passive base region; in this region, eq. (8) holds. Around the passive base region is the boundary of the base diffusion. A boundary condition is that no current flows through this boundary; the current flowing through the boundary is proportional to the normally-directed derivative of the potential, so on this boundary

$$\mathbf{n}\cdot\nabla\varphi \equiv \frac{\partial\varphi}{\partial n} = 0, \tag{13}$$

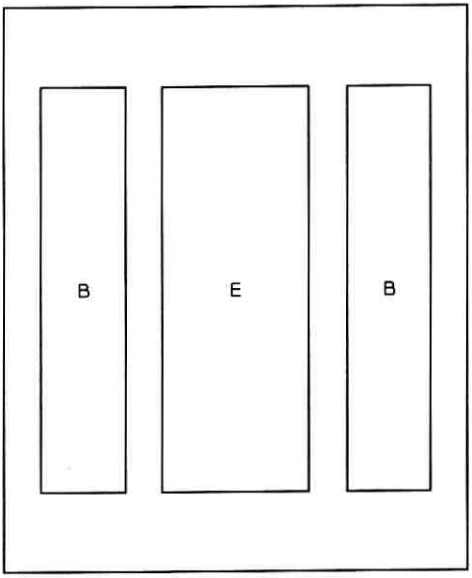

Fig. 3—Top view of transistor used in calculations.

where $\mathbf{n}$ is the outer-directed unit normal to the boundary. At the edge of the emitter diffusion, the dividing line between the active and passive base regions, the boundary conditions are that the potential be continuous and that the total current flow be continuous. These conditions are

$$\varphi_A = \varphi_P, \tag{14}$$

$$\mathbf{n}\cdot\left(\frac{1}{R_A}\nabla\varphi_A\right) = \mathbf{n}\cdot\left(\frac{1}{R_P}\nabla\varphi_P\right), \tag{15}$$

where $\mathbf{n}$ is the unit normal outward from the emitter diffusion, φ_A refers to the potential just inside the active base, and φ_P refers to the potential just inside the passive base. Equation (14) neglects the edge injection of current from the emitter-base junction; such injection could easily be included. It leads to a boundary condition, similar to eq. (15), but involving both φ and $\partial\varphi/\partial n$.

III. SOLUTION TECHNIQUE

This section describes an economical method for solving the coupled partial differential equations (8) and (10) with boundary conditions (13) through (15). The usual way to solve such equations is by finite dif-

ferences. A square or rectangular grid is superimposed on the x, y plane and derivatives at the grid vertices are approximated by finite differences. A large number of linear algebraic equations are generated and solved iteratively. While very powerful, finite difference methods have several disadvantages for the problem at hand.

It may not be easy to get the transistor boundaries (emitter diffusion, base contacts, and base diffusion) to fall on grid lines, so that more complicated programming is necessary. The areas in which the equations must be solved are frequently odd-shaped (see Fig. 4). In such cases, the usual iteration schemes are not guaranteed to converge and it may be necessary to develop an efficient scheme.

The method[3,4] developed for solving equations like (11) does not employ a grid and does not require iteration to solve the set of linear algebraic equations. The method will be briefly described for a single equation,

$$\nabla^2\varphi = \gamma^2\varphi, \tag{16}$$

holding in some area A with boundary Γ. The area A may be any shape; it may even be multiply connected. For (16), Ref. 4 derives the exact expression,

$$\pi\varphi(s') = P\int_\Gamma \left[K_0(\gamma r)\frac{\partial\varphi(s)}{\partial n} + (\mathbf{n}\cdot\mathbf{r}/r)K_1(\gamma r)\varphi(s)\right] ds. \tag{17}$$

Here s' is any point on the boundary (except at a corner point), P denotes the Cauchy principle value, K_0 and K_1 are modified Bessel functions of the first kind,[5] and $\mathbf{r}$ is the vector from s' to s. Points inside the region A are not involved; however, if s' is inside A, (17) can be

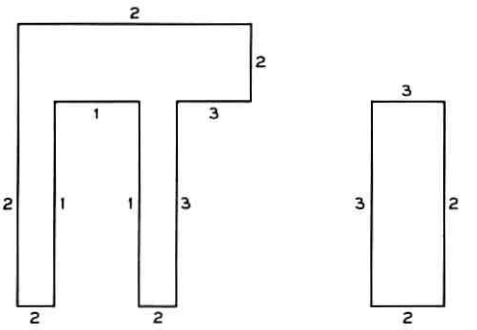

Fig. 4—"Exploded" view of one-fourth of transistor. Left—passive base region; right—active base region (under emitter). Numbers indicate boundary conditions applying (see text).

used if π is replaced by 2π. Thus, if φ and $\partial\varphi/\partial n$ are known on the boundary, φ may be found inside the boundary.

To find $\varphi(s)$ and $\partial\varphi(s)/\partial n$ on the boundary, (17) is regarded as an integral equation. This and the boundary conditions for the original partial differential equation suffice to determine the solution. In principle, an accurate numerical solution is quite simple, although the computer programming is rather complex. Various methods may be used to solve (17) plus the boundary conditions; the method used in Ref. 4 will be described here. The boundary Γ is divided into a number of straight-line segments; on each segment $\varphi(s)$ and $\partial\varphi(s)/\partial n$ are approximated by polynomials of any desired order with unknown coefficients. Using geometrical information about Γ, and the expansion of K_0 and K_1 in terms of logarithms and polynomials,[5] the integrals in (17) may be done symbolically.[4] Equation (17) cannot in general be satisfied exactly at each point s' along the boundary; N points are chosen at which to satisfy (17) exactly, where N is the total number of unknown coefficients in all of the polynomials approximating φ and $\partial\varphi/\partial n$, on every part of Γ. Then (17) is replaced by N linear algebraic equations for the polynomial coefficients. For typical transistors, N is no more than 50 or so, and therefore the equations may be solved directly without iteration by Gaussian elimination; $N = 50$ gives an accuracy of about one percent for the numerical example considered later.

When the polynomial coefficients have been found, φ and $\partial\varphi/\partial n$ are known on the boundary and φ may be found inside A if it is desired. For the transistor analysis, however, φ is not needed inside A; it is only needed on the boundary. Although the program is quite complex, the computer time is quite short. For Laplace's equation, the boundary integral equation method is considerably faster than finite difference methods.[4]

The two coupled partial differential equations, (8) and (10), may be solved in a similar manner. Fig. 4 shows the areas of Fig. 3 in which each of the two equations can be used. Because of the fourfold symmetry of Fig. 3, only one-fourth of the transistor need be considered, and only one-fourth is shown. On the left is the passive base region, in which (10) holds; on the right is the active base region, in which (8) holds. The numbers on the line segments indicate the boundary conditions on each segment. On segments numbered 1, the base contacts, $(\varphi - v_{CE})$ is constant; on the exterior of the base diffusion, segments numbered 2, and symmetry lines, $\partial\varphi/\partial n = 0$; and on segments numbered 3, the boundary of the passive and active base regions, (14) and (15) hold.

IV. CALCULATION OF h-PARAMETERS

With a fast and accurate method of solving the coupled partial differential equations (8) and (10), finding the small-signal parameters of the transistor is relatively easy. This section describes the method for determining common-emitter h-parameters.

Common-emitter h-parameters are defined by

$$v_{BE} = h_{11}i_B + h_{12}v_{CE} = h_{ie}i_B + h_{re}v_{CE}, \tag{18a}$$

$$i_C = h_{21}i_B + h_{22}v_{CE} = h_{fe}i_B + h_{oe}v_{CE}. \tag{18b}$$

In order to find all four h-parameters at a given frequency, two potential solutions are necessary. The first solution gives h_{11} and h_{21} and assumes v_{CE} to be zero. With $v_{CE} = 0$, $v_{BE} = h_{11}i_B$ and $i_C = h_{21}i_B$. The base contacts are taken to be at unit potential, $v_{BE} = 1$. The boundary integral equation method is used to solve (8) and (10) with boundary conditions (13), (14), (15), plus $\varphi = 1$ on the base contact (lines numbered 1 in Fig. 4). In Fig. 4, the output is $\partial\varphi/\partial n$ on lines numbered 1, where $\varphi = 1$; φ on lines numbered 2, where $\partial\varphi/\partial n = 0$; and both φ and $\partial\varphi/\partial n$ on lines numbered 3. The current flowing into the base contact is just

$$i_B = \frac{1}{R_P}\int \frac{\partial\varphi(s)}{\partial n}\,ds, \tag{19}$$

where the integration is over lines numbered 1 in Fig. 4. Since $\partial\varphi/\partial n$ is approximated by a polynomial on each of the sides, the integral in (19) is trivial. This gives $h_{11} = 1/i_B$.

To find h_{21}, the collector current must be found, which involves the modeling of the transistor in the z-direction. Again, the simple model used may be interpreted with the aid of Fig. 2. Transit-time effects in base and collector will be neglected. The z-directed collector current flowing through a small area $\Delta x\ \Delta y$ has three components:

(*i*) Diffusion of injected charge across the active base region. G_m is a conductance per unit area.

$$(\Delta i_C)_1 = G_m\varphi\ \Delta x\ \Delta y. \tag{20}$$

(*ii*) Injected current from collector-base junction capacitance.

$$(\Delta i_C)_2 = -C_C\frac{\partial}{\partial t}(\varphi - v_{CE})\ \Delta x\ \Delta y. \tag{21}$$

(*iii*) Injected current from base-width modulation in the active base region.

$$(\Delta i_C)_3 = G_0 v_{CE}\,\Delta x\,\Delta y - \left(G' + C'\frac{\partial}{\partial t}\right)(\varphi - v_{CE})\,\Delta x\,\Delta y. \tag{22}$$

From the passive base region, current flows to the collector only through the capacitance C_C .

The total collector current is

$$\begin{aligned} i_C = \iint_{\text{Active Base}} &\{[(G_m - G') - i\omega(C_C + C')]\varphi \\ &+ [(G_0 + G') + i\omega(C_C + C')]v_{CE}\}\,dx\,dy \\ &- \iint_{\text{Passive Base}} i\omega_C(\varphi - v_{CE})\,dx\,dy. \end{aligned} \tag{23}$$

In the calculations, G' was neglected, $G_0 + G' \equiv G_1$ were combined, and C' was neglected.

Thus it is necessary to do an area integral of the potential. Since the factors multiplying φ are independent of x and y, a two-dimensional numerical integration may be avoided. Consider one of Green's boundary-value formulas,

$$\iint_A (G\nabla^2\varphi - \varphi\nabla^2 G)\,dx\,dy = \oint_\Gamma \left(G\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial G}{\partial n}\right) ds, \tag{24}$$

where the line integral is over Γ, the boundary of the area A. This is true for any well-behaved functions $\varphi(x, y)$ and $G(x, y)$. Suppose that φ obeys (16). In the area integral, substitute $\gamma^2\varphi$ for $\nabla^2\varphi$ to obtain

$$-\iint_A \varphi(\nabla^2 G - \gamma^2 G)\,dx\,dy = \oint_\Gamma \left(G\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial G}{\partial n}\right) ds. \tag{25}$$

Now G may be chosen so that

$$\nabla^2 G - \gamma^2 G = 1, \tag{26}$$

or

$$G = -1/\gamma^2,$$

to give

$$\iint_A \varphi\,dx\,dy = -\frac{1}{\gamma^2}\oint_\Gamma \frac{\partial\varphi}{\partial n}\,ds. \tag{27}$$

Thus the area integral in (23) may be replaced by a line integral;

since $\partial\varphi/\partial n$ is a known polynomial, the integral is again trivial. In this way i_C may be found, and $h_{21} = i_C/i_B$.

In order to find h_{12} and h_{22} , i_B is assumed to be zero. With $i_B = 0$, $v_{BE} = h_{12}v_{CE}$ and $i_C = h_{22}v_{CE}$. The potential at the base contacts, v_{BE} , is unknown; since $i_B = 0$, $\partial\varphi/\partial n = 0$ at the base. For the second solution, the boundary integral equation method is used to solve (8) and (10) for $(\varphi - v_{CE})$ with boundary conditions (13), (14), (15), plus $\varphi = v_{BE} =$ unknown constant and $\partial\varphi/\partial n = 0$ on the base contact (lines numbered 1 in Fig. 4). v_{CE} is assumed to be 1. The output is φ on lines numbered 1 and 2, where $\partial\varphi/\partial n = 0$, and both φ and $\partial\varphi/\partial n$ on lines numbered 3. Then $v_{BE} = h_{12}$ is found directly from φ on the base contact, and i_C may be found as before to give $h_{22} = i_C$.

V. RESULTS FROM SAMPLE CALCULATION

Experimental h-parameters were available for a double-diffused, double-base-stripe silicon transistor, similar to that shown in Fig. 3. The data were taken with $v_{CE} = 1$ V, and $I_E = 10$ mA (approximately 230 A/cm^2). In the theoretical calculations, there are seven parameters: R_P , C_C , R_A , G_m , G_E , C_E , and G_1 . Two of these, R_P and G_m , were not adjusted. R_P , the sheet resistance of the passive base, was given its nominal fabrication value. G_m , which is essentially a transconductance per unit area, was set equal to $I_E/(A_E kT/q)$. A_E is the emitter area. The other five parameters were chosen so that the experimental and calculated h-parameter matched at low frequencies (below 1 MHz). No additional adjustments were made to fit the high-frequency data. The parameter values used in the calculation are given in Table I. For the calculations, the nominal fabrication geometry was used.

Complex-plane plots of the four common-emitter h-parameters are given in Figs. 5 through 8. The solid lines are drawn from the experimental data, only a few points of which are shown (solid circles). Fre-

TABLE I—PARAMETERS USED IN FITTING h-PARAMETERS

R_P,	sheet resistance in passive base, 170 ohms/□
R_A,	sheet resistance in active base, 6000 ohms/□
G_E,	conductance per unit area of emitter-base junction, 0.00051 mho/mil^2 = 79 mho/cm^2
C_E,	capacitance per unit area of emitter-base junction, 21 pF/mil^2 = 330 μF/cm^2
G_1,	low-frequency limit of h_{oc}, 0.0003 mho
C_C,	capacitance per unit area of collector-base junction, 0.226 pF/mil^2 = 0.0350 μF/cm^2
G_m,	"transconductance" per unit area, 0.058 mho/mil^2 = 8900 mho/cm^2

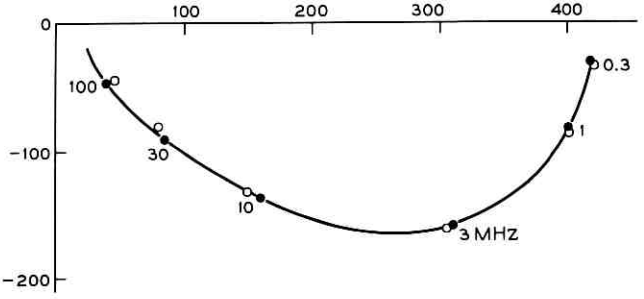

Fig. 5—Input impedance, h_{11} or h_{ie}, in ohms. The solid line and circles are experimental, and the open circles theoretical. Frequency is a parameter along the curve, and selected frequencies are given in MHz.

quency is a parameter along each curve; selected frequencies are indicated in MHz. At these frequencies, the calculated h-parameters are plotted as open circles.

At the higher frequencies, the discrepancy between experimental and calculated results becomes significant. There are several reasons why this is to be expected. In the first place, the experimental data may be in error, since h-parameter measurements are considerably more difficult to obtain at high than at low frequencies. However, it is not necessary to invoke this explanation (the traditional one for theorists) for the discrepancies to be understood.

There are two other sources of discrepancy. The first is that the simple model used for the transistor is not adequate at high frequencies. "Second-order" effects such as transit-time delay in the base and collector were ignored; other effects which are small at low frequencies could also be important. At high frequencies, the exact shape and size

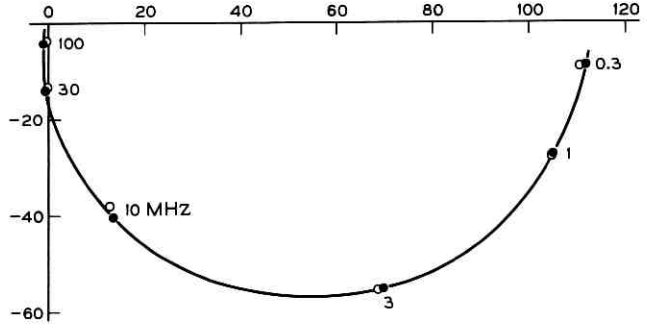

Fig. 6—Forward current gain, h_{21} or h_{fe}, dimensionless.

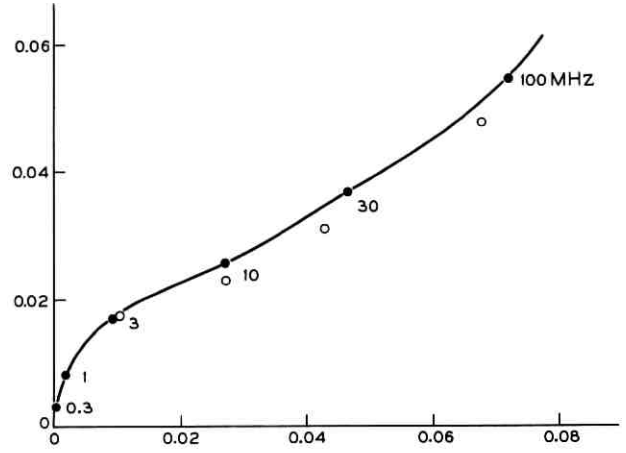

Fig. 7—Reverse voltage transfer factor, h_{12} or h_{re}, dimensionless.

of the transistor is important and the measured transistor may certainly differ somewhat in geometry from the nominal. The effect of geometry may be seen in the plot for h_{22} . The loop in h_{22} at high frequencies is due to the geometry of the transistor. Calculations done with other geometries show that, as the base and emitter stripes are made longer, the horizontal width of the loop shrinks; the loop is absent for a transistor with infinitely long stripes.

Secondly, at high frequencies the parasitic effects of the header and the encapsulation become important and need to be accurately modeled if the experimental data are to be matched. Sample calculations with

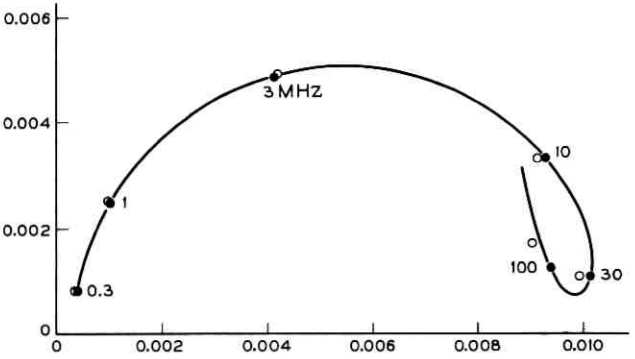

Fig. 8—Output conductance, h_{22} or h_{oe}, in mhos.

representative values showed that header parasitics had at least as large an effect on the h-parameters as the high-frequency discrepancy.

REFERENCES

1. Gwyn, C. N., Scharfetter, D. L., and Wirth, J. L., "The Analysis of Radiation Effects in Semiconductor Junction Devices," IEEE Trans. Nucl. Sci., *NS-14*, (December 1967), pp. 153–169.
2. Gummel, H. K., and Poon, H. C., "An Integral Charge Control Model of Bipolar Transistors," B.S.T.J., *49*, No. 5 (May–June 1970), pp. 827–846.
3. Blue, J. L., "Two-Dimensional Distributed Base Resistance Effects in Bipolar Transistors," talk presented at the 1969 International Electron Device Meeting, October 31, 1969, Washington, D. C.
4. Blue, J. L., "On Finding the Admittance Matrix of a Thin-Film Network by Solving the Reduced Wave Equation in Two Dimensions," J. Comp. Phys., *7*, (April 1971), pp. 327–345.
5. Abramowitz, M., and Stegun, I., Eds., *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards, 1964, pp. 374–379.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Stability of Distributed Systems With Feedback via Michailov's Criterion

By G. C. REIS*

(Manuscript received November 11, 1971)

This paper is based on results derived during a stability study of the Saturn V rocket for which it was necessary to validate the use of Nyquist's encirclement-counting technique in distributed systems. An outline of the paper is as follows: Certain results concerning the finiteness of the number of zeros of polynomials in s and e^s are shown in Theorem 1 and its corollaries. Theorem 2 is a generalization of Michailov's Criterion. Simplifying assumptions, usually valid in practice, yield a simplified test to determine if "encirclement-counting" is a valid stability test [equation (20)]. The results are reformulated for an open-loop analysis. Various aspects of the theory are shown by three examples based on an electrical equivalent of a simple single-engine, liquid-fuel rocket.

I. INTRODUCTION

Liquid-fueled rockets can exhibit a peculiar type of instability due to self-sustained longitudinal oscillations. Since the rocket then stretches and shrinks longitudinally, it behaves like a pogo-stick, which has resulted in the nickname POGO for this type of instability.

To see how this phenomenon arises, consider the simple diagram shown in Fig. 1. The chain of events which can cause POGO is initiated by a random variation in thrust of the engine. This thrust variation causes the rocket structure to oscillate in its natural modes. The pressure in the fuel tank thus varies. This pressure variation is propagated down the fuel feed line, resulting in a variation of fuel flow into the engine. Since the thrust of the engine is proportional to the rate of fuel entering, the loop is completed, and instability results if this resulting thrust variation aids the original random variation which initiated the chain of events.

* This work was done when the author was with Bellcomm, Inc., and therefore was performed under NASA contract NSW417.

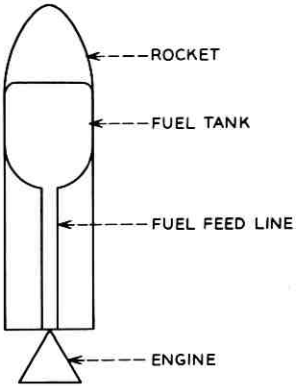

Fig. 1—Simple model of a liquid-fuel rocket.

In modeling this physical situation, it is customary to perform a modal analysis of the rocket structure, and retain only the most significant modes. This results in the distributed structure being replaced by a lumped approximation. Although the same technique could be used to lump the feed-line, there are many reasons for desiring to keep this element as a distributed parameter. There is no great difficulty in doing so, since the fluid equations which govern the feed-line are of the same form as electrical transmission lines.

An equivalent circuit of Fig. 1 would then be Fig. 2, where $V_1(s)$, $V_2(s)$ are the Laplace transforms of the pressure variations at the top and bottom of the feed-line, respectively, and $I_1(s)$, $I_2(s)$ represent the transform of flow variations. $E_1(s)$ is then the random pressure variation at the top of the line due to the assumed random thrust variation above. $Y(s)$ represents the hydro-mechanical impedance of the feed-line output, and $G(s)$ includes the structural feedback. To see the type of equations which will be of concern for a stability analysis, assume that losses can be neglected in the feed-line.

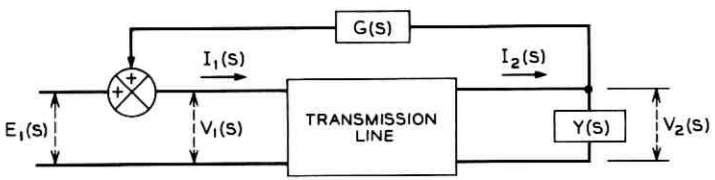

Fig. 2—Electrical equivalent of Fig. 1.

The system equations can be written in matrix form as:

$$\begin{bmatrix} \cosh ks & -Z_c \sinh ks & -1 & 0 \\ -\frac{1}{Z_c}\sinh ks & \cosh ks & 0 & -1 \\ -1 & 0 & G(s) & 0 \\ 0 & 0 & Y(s) & -1 \end{bmatrix} \begin{bmatrix} V_1(s) \\ I_1(s) \\ V_2(s) \\ I_2(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -E_1(s) \\ 0 \end{bmatrix} \tag{1}$$

where Z_c (a positive real number) is the characteristic impedance of the line and k (a positive real number) is the ratio of line length to wave speed in the line. The determinant ($\Delta(s)$) of the matrix appearing in (1) is of interest in a stability analysis. It can easily be computed to be

$$\Delta(s) = \cosh ks - G(s) + Z_c Y(s) \sinh ks. \tag{2}$$

In general, neither $Y(s)$ nor $G(s)$ need be rational (especially in the case of multiple engines) but for simplicity, suppose that

$$G(s) = \frac{N_G(s)}{D_G(s)} \qquad Z_c Y(s) = \frac{N_Y(s)}{D_Y(s)} \tag{3}$$

where N_G , D_G , N_Y , D_Y are polynomials in the complex variable s. Then (2) can be written as

$$\Delta(s)[D_G(s)D_Y(s)] = [D_G(s)D_Y(s)] \cosh ks - N_G(s)D_Y(s) + D_G(s)N_Y(s) \sinh ks. \tag{4}$$

In the following sections of this paper we will be concerned with polynomials in the two complex variable s and e^s. We now show that (4) can be considered as such. Thus multiply (4) by $2e^{-2ks}$ to give

$$F(s) = 2e^{-ks}\,\Delta(s)[D_G(s)\,D_Y(s)] = D_G(s)[D_Y(s) + N_Y(s)] - 2N_G(s)\,D_Y(s)e^{-ks} + D_G(s)[D_Y(s) - N_Y(s)]e^{-2ks} \tag{5}$$

or

$$F(s) = \sum_{i=0}^{2} R_i(s)e^{-iks} \tag{6}$$

where each $R_i(s)$ is a polynomial in s.

This paper will also be concerned with open-loop and closed-loop expressions, and will assume that the quantities of interest will be of the form of (6). To show that this is true in our present example, assume that $G(s)$ "closes the loop" and solve for the "line transfer function"

$V_2(s)/V_1(s)$. The result is

$$\frac{V_2(s)}{V_1(s)} = \frac{1}{\cosh ks + Z_C Y(s) \sinh ks}. \tag{7}$$

Hence the "loop gain," $G_0(s)$, is simply $G(s)$ times (7), or

$$G_0(s) = \frac{G(s)}{\cosh ks + Z_C Y(s) \sinh ks}$$

$$= \frac{D_Y(s)N_G(s)}{D_Y(s)\, D_G(s) \cosh ks + D_G(s)N_Y(s) \sinh ks}. \tag{8}$$

The denominator of (8) is seen to be of the form of $\Delta(s)$, and hence can be made to look like (6).

Finally, we wish to remove the simplifying assumption that $Y(s)$ and $G(s)$ are rational. This is desired since an actual engine is not only fed liquid fuel, but also liquid oxidizer. Thus even a single-engine rocket has two feed-lines. (The Saturn V has five engines, for a total of ten feed-lines.) The structure of a model for rockets of this complexity would be that of Fig. 2, repeated once for each feed-line, with suitable interconnections through lumped (i.e., rational) transfer functions. Thus the $G(s)$ and $Y(s)$ of Fig. 2 will be the ratio of sums of powers of s and e^s. As such they can be combined to yield forms such as (6).

A final constraint on the stability analysis is that it is required that the analysis be of the conventional open-loop type using Nyquist's Criterion. For lumped systems this presents little difficulty since one can always make open-loop measurements at as high a frequency as necessary to guarantee that all singularities of the transfer function are included. Furthermore, for rational functions, no difficulty is encountered in closing the contour in the right-half s-plane. With distributed systems, however, it is possible that the gain becomes periodic for large magnitude of s, and care must be exercised in determining closed-loop stability via open-loop gain plots. What is desired, therefore, is a set of conditions under which the conventional "encirclement counting" technique for lumped systems, remains valid for distributed systems of the type described.

There are two important techniques for determining whether a polynomial in the two complex variables s and e^s has any zeros for Re $[s] \geqq 0$. These are the Pontryagin Criterion[1] and the Michailov Criterion.[2] It is of interest to see if these criteria can be applied to the ratio of such polynomials in order to determine stability of closed-loop gain. In a previous note[3] it was shown that this is not feasible for criteria

of the Pontryagin type. In this paper we are able to develop a stability criterion of the Nyquist type from Michailov's Criterion for a large class of distributed-parameter systems, in particular, for a large class of transmission line systems with feedback.

An outline of the paper is as follows: Certain results concerning the finiteness of the number of zeros of polynomials in s and e^s are stated as Theorem 1 and its corollaries. Theorem 2 is the desired generalization of Michailov's Criterion. Simplifying assumptions, usually valid in practice, yield a simplified test to determine if "encirclement-counting" is a valid stability test [equation (20)]. The results are reformulated for an open-loop analysis. Three examples show various aspects of the theoretical analysis. The Appendix includes a statement of Pontryagin's Criterion suitable for use in the present paper. Also included are the proofs of the two theorems and a derivation of some conditions under which Michailov's Criterion can be simplified.

II. MICHAILOV'S CRITERION

As a starting point we consider an equation of the form

$$G(z) = \sum_{i=1}^{m} \sum_{j=0}^{n} \bar{a}_{ij} z^j e^{\omega_i z} = 0 \tag{9}$$

where $\bar{a}_{ij}$ are complex and ω_i are real. If any of the ω_i were negative, we could multiply $G(z)$ by $e^{|\omega_k| z}$, where ω_k is the most negative of the ωs. This would not change the zeros of $G(z)$, so we assume $0 \leqq \omega_1 < \omega_2 < \cdots < \omega_m$. Dividing by $e^{\omega_m z}$ and letting $\bar{a}_{m-i+1,j} = a_{ij}$, we transform (9) into

$$F(z) = \sum_{i=1}^{m} \sum_{j=0}^{n} a_{ij} z^j e^{-r_i z} \tag{10}$$

where $r_i = \omega_m - \omega_{m-i+1} > 0$ for $i = 2, 3, \cdots, m$ and $r_1 = 0$. [To relate this to the Pontryagin Criterion, note that if the ω_i are rational (which can always be assumed in a practical situation) then a suitable scaling of the z variable will make $G(z)$ of (9) into a polynomial like $H(z)$ of the Pontryagin Criterion.]*

Before continuing, it should be noted that a proof of the Michailov Criterion for exponential polynomials has been presented in the literature.[2] However, this proof assumed that

$$| a_{1n} | > \sum_{i=2}^{m} | a_{in} |. \tag{11}$$

* See Appendix and Ref. 1.

As will be seen by the example to be considered later, in transmission line systems, (11) is almost never satisfied. Hence, a proof is desired which is free of this assumption. However, we do use the assumption that $a_{1n} \neq 0$. That this is no loss of generality can be seen by the following considerations. It is clear that $| r_m | \geqq | r_i |$, $i = 2, \cdots, m$. Then multiplying (10) by $e^{r_m z}$ (which does not change the zeros) puts (10) into the Pontryagin form. If a_{1n} is zero, the principal term is missing and we are finished with the stability study.* To aid in the subsequent development, let us rewrite (10) as

$$F(z) = \sum_{i=0}^{n} z^i Q_i(e^{-z}) \tag{12}$$

where $Q_i(e^{-z}) = \sum_{i=1}^{m} a_{ij} e^{-r_i z}$, or as

$$F(z) = z^k F_k(z) + \sum_{i=0}^{k-1} z^i Q_i(e^{-z}) \tag{13}$$

where k is an integer between zero and n and

$$F_k(z) = \sum_{i=k}^{n} z^{i-k} Q_i(e^{-z}). \tag{14}$$

In the Appendix we prove the following theorem:

Theorem 1: If there exists a non-negative integer $k \leqq n$ such that $F_k(z)$ of (14) has at most a finite number of zeros on Re $(z) \geqq 0$, *then $F(z)$ of (13) has at most a finite number of zeros on* Re $(z) > 0$.

The following corollaries are of interest:

Corollary 1: If there exists a non-negative integer $k \leqq n$ such that $F_k(z)$ of (14) has no zeros on Re $(z) \geqq 0$, *then $F(z)$ of (12) has at most a finite number of zeros on* Re $(z) > 0$.

Corollary 2: If $Q_n(e^{-z})$ of (12) has no zeros on Re $(z) \geqq 0$ *then $F(z)$ of (12) has at most a finite number of zeros on* Re $(z) > 0$.

[Corollary 2 follows since $Q_n(e^{-z}) = F_n(z)$.]

Let

$$F(z) = 1 + \psi(z) z^k F_k(z) \tag{15}$$

where

$$\psi(z) = \sum_{i=0}^{k-1} z^{(i-k)} \frac{Q_i(e^{-z})}{F_k(z)}. \tag{16}$$

In the Appendix we derive theorem 2:

* See Appendix and Ref. 1.

Theorem 2: The number of zeros of $F(z)$ with positive real part is

$$N = \frac{k}{2} - \frac{1}{2\pi}\Delta_{-\omega}(F(z)) + \frac{\Delta_C}{2\pi}(F_k(z)) + \frac{1}{2\pi}\arg(1 + \psi(iy)) - \arg(1 + \psi(-iy)) \quad (17)$$

assuming that $F(z)$ has no purely imaginary zeros and where $\Delta_{-\omega}(F(z))$ is the net change in arg $F(z)$ *along the imaginary axis from $-iy$ to $+iy$, and $\Delta_C(F_k(z))$ is the net change in* arg $(F_k(z))$ *along contuor C, any contour outside the semicircle of radius R of Theorem 1.*

III. MICHAILOV'S CRITERION: SPECIAL CASE

Theorem 2 is the desired statement of Michailov's Criterion. To obtain tighter results, let us now assume that $Q_n(e^{-z})$ has no zeros on Re $(z) \geqq 0$ (i.e., Corollary 2). Further, let the r_i be rational and the a_{ij} be real. By virtue of rational r_i, $Q_k(e^{-z})$ is periodic in y, for $z = x + iy$. Let this period be P. By virtue of real a_{ij}, replacing z by its conjugate results in $Q_k(e^{-z})$ being replaced by its conjugate. Hence we need only consider the semi-infinite strip defined by $x > 0$ and $P \geqq y \geqq 0$.

Michailov's Criterion can be simplified if it can be shown that $Q_n(e^{-z})$ does not wind around the origin as z varies over a suitable C. We now consider this possibility.

$$\begin{aligned} Q_n(e^{-z}) &= \sum_{i=1}^{m} a_{in}e^{-r_i z}(\cos r_i y - i \sin r_i y) \\ &= a_{1n} + \sum_{i=2}^{m} a_{in}e^{-r_i z}\cos r_i y - i\sum_{i=2}^{m} a_{in}e^{-r_i z}\sin r_i y \\ &= \text{Re}\,[Q_n] - i\,\text{Im}\,[Q_n]. \end{aligned} \quad (18)$$

If either Re $[Q_n]$ or Im $[Q_n]$ does not vanish along C, then Q_n cannot wind around the origin. In the Appendix we derive sufficient conditions for this.

We now assume that $\Delta_C[Q_n(e^{-z})] = 0$ and write the Michailov Criterion as

$$N = \frac{n}{2} - \frac{1}{\pi}\Delta_{-w/2}(F(z)) + \frac{1}{\pi}\arg(1 + \psi(iy)) \quad (19)$$

where $\Delta_{-w/2}$ is that part of the imaginary axis from 0 to iy. Thus $N = 0$ if

and only if

$$\Delta_{-w/2}(F(z)) = \frac{n\pi}{2} + \arg\,(1 + \psi(iy)). \tag{20}$$

We remark that arg $(1 + \psi(iy))$ can be made close to zero for y sufficiently large.

IV. MICHAILOV'S CRITERION APPLIED TO OPEN-LOOP ANALYSIS

The Michailov Criterion, as well as its predecessor the Pontryagin Criterion, settle the problem of finding rhp zeros of polynomials in z and e^z. In many engineering applications, however, this polynomial is not directly available, but a related ratio of such polynomials can be found. In the study of feedback systems, for example, an open-loop gain can be measured and it is desired to find the poles of the closed-loop gain. These latter poles are the zeros of the polynomial which results from adding the two polynomials whose ratio is the open-loop gain. Stability has been determined for nondistributed systems by counting encirclements of the open-loop gain along the imaginary axis. What we propose to do next is to provide a similar criterion for the distributed parameter problem. Thus let $F(z) = D(z) + N(z)$. Then

$$\begin{aligned}\Delta_\Gamma(D(z) + N(z)) &= \Delta_\Gamma(D(z))\left(1 + \frac{N(z)}{D(z)}\right)\\ &= \Delta_\Gamma(D(z)) + \Delta_\Gamma\left(1 + \frac{N(z)}{D(z)}\right).\end{aligned} \tag{21}$$

Let the contour Γ be composed of a portion of the imaginary axis w and another (possibly semicircular) contour C, such that Γ encloses all zeros of $D(z) + N(z)$. Then

$$\Delta_\Gamma(D(z) + N(z)) = \Delta_\Gamma(D(z)) + \Delta_w\left(1 + \frac{N(z)}{D(z)}\right) + \Delta_C\left(1 + \frac{N(z)}{D(z)}\right). \tag{22}$$

It is the term

$$\Delta_w\left(1 + \frac{N(z)}{D(z)}\right)$$

which is usually available for determining stability. We ask,

"When does $\Delta_w\left(1 + \frac{N(z)}{D(z)}\right) = \Delta_\Gamma(D(z) + N(z))$?".

The answer is that this happens exactly when

$$0 = \Delta_\Gamma(D(z)) + \Delta_C\left(1 + \frac{N(z)}{D(z)}\right). \tag{23}$$

To develop a more practical criterion let us rewrite this expression using n_F to be the highest power of z in $F(z)$, and N_F to be the number of zeros of $F(z)$ inside Γ. Then

$$0 = \Delta_\Gamma(D(z)) + \Delta_C(D(z) + N(z)) - \Delta_C(D(z)) \tag{24}$$

$$0 = 2\pi N_D + n_{D+N}\pi - n_D\pi \tag{25}$$

where we have neglected those terms which become small for large z. If we limit further consideration to systems which are open-loop stable (i.e., $N_D = 0$) then (25) requires that $n_{D+N} = n_D$. In most practical situations, the open-loop gain is bounded at infinity, that is to say $n_N \leqq n_D$. Hence $n_{D+N} \leqq n_D$. Since $n_{D+N} < n_D$ requires

$$\lim_{z\to\infty} \frac{N(z)}{D(z)} = -1,$$

we can conclude that counting encirclements of the open-loop gain is a valid method for determining stability [i.e., (25) is satisfied] for systems which are open-loop stable ($N_D = 0$) and whose gain is bounded at infinity ($n_{D+N} \leqq n_D$) but does not approach -1 for large frequencies ($n_{D+N} \not< n_D$). This includes the case, usually found in practice, that the open-loop gain approaches zero for large frequencies.

V. EXAMPLES

All examples refer to Fig. 2 and are chosen to illustrate various aspects of the analysis. At various points in the examples, the following assumptions concerning $G(s)$ and $Y(s)$ are referred to:

A1. $G(s)$ and $Y(s)$ are each the ratio of two polynomials having real coefficients with no singularities on Re $(s) > 0$.

A2. $\lim_{s\to\infty} G(s) = k_1$ where k_1 is real and $|k_1| \leqq 1$.

A3. $\lim_{s\to\infty} Y(s) = \lim_{s\to\infty} sC$ where C is non-negative real.

Assumption A1 requires G and Y to be stable transfer functions. Assumption A2 insists that the feedback gain at infinity be less than unity. Assumption A3 is physically appealing.

Some unusual properties of the natural frequencies of this system have been described elsewhere.[4,5]

Example 1: In this example, Assumptions A1, A2, and A3 are invoked, and $G(s)$ and $Y(s)$ are given by (3). We wish to show that assumption

(11), used in previous derivations of Michailov's Criterion, is not met and that it is valid to count encirclements of the Nyquist plot to determine stability. By A1, $D_G(s)$ and $D_Y(s)$ have no right-half plane zeros. Thus $\Delta(s)$ has right-half plane zeros exactly when $\Delta(s)D_G(s)D_Y(s)$ has right-half plane zeros.

From Assumption A2 we conclude that deg $D_G(s) \geqq$ deg $N_G(s)$. From A3 we conclude that deg $N_Y(s) >$ deg $D_Y(s)$. From this, and (5) and (6), we see that the principal term is present and that the assumption (11) used in previous proofs of Michailov's Criterion is not met. In fact, one can readily convince oneself that this will be the case whenever lossless transmission lines are involved, since all exponential terms will involve hyperbolic functions.

Using the notation of (6), the open-loop gain (8) can be expressed as

$$G_0(s) = \frac{R_1(s)e^{-ks}}{R_0(s) + R_2(s)e^{-2ks}} \tag{26}$$

whose norm becomes small for large, right-half plane values of s. Hence it is valid to count encirclements of the open-loop gain about the point $+1$.

Example 2: Here we show the necessity of the open-loop stability requirement. Suppose Assumptions A1 and A2 are invoked and further assume that $D_G(s) = 1$, $N_G(s) = k_1$, $D_Y(s) = 1$, $N_Y(s) = sC_1 + g$. This corresponds to terminating the line in a capacitance C_1 and shunt conductance $g \neq 0$. The open-loop gain for $s = jw$ becomes

$$G_0(jw) = \frac{k_1}{\cos kw - wC_1 \sin kw + ig \sin kw} \tag{27}$$

which is real only when $\sin kw$ is zero. This implies that $\cos kw$ is ± 1. Thus if $| k_1 | < 1$, $G_0(jw)$ cannot encircle the $+1$ point. (This result is in agreement with Assertion 2 of Ref. 5, to which this problem corresponds if $g = 0$. It is intuitive that adding losses to a lossless system will enhance stability.)

To show the necessity of the open-loop stable requirement, note that g can be either positive or negative. From the Pontryagin Criterion, we see that (27) is then stable or unstable, respectively, and that for $| k_1 | < 1$ the closed-loop system is stable or unstable, respectively. However, in either case there are no encirclements of the critical point by the open-loop gain.

Example 3: Let

$$G(s) = \frac{\sum_{i=0}^{n} a_i s^i}{\sum_{j=0}^{n} b_j s^j} \qquad b_n \neq 0$$

$$Z_C Y(s) = \frac{\sum_{k=0}^{p} c_k s^k}{\sum_{l=0}^{p} d_l s^l} \qquad c_p \neq 0. \tag{28}$$

In this final example we look at how the a, b, c, and d coefficients of (28) enter into the $R_i(s)$ polynomials of (6) and into the $Q_n(e^{-z})$ polynomials of Theorem 1 and its corollaries. We show how the Assumptions A2 and A3 affect whether the system satisfies Corollaries 1 and 2, and thereby provide examples of such systems. Using (3) and (28), (5) becomes

$$\begin{aligned} F(s) = {} & s^{n+p}\{b_n(d_p + c_p) - 2a_n d_p e^{-ks} + b_n(d_p - c_p)e^{-2ks}\} \\ & + s^{n+p-1}\{[(d_{p-1} + c_{p-1})b_n + b_{n-1}(d_p + c_p) \\ & - 2[a_{n-1}d_p + a_n d_{p-1}]e^{-ks} \\ & + [b_n(d_{p-1} - c_{p-1}) + b_{n-1}(d_p - c_p)]e^{-2ks}\} \\ & + \sum_{m=0}^{n+p-2} s^m\{\sum_{i+i=m} b_i(d_i + c_i) - 2e^{-ks}\sum_{i+i=m} a_i d_i \\ & + e^{-2ks}\sum_{j+i=m} b_j(d_i - c_i)\}. \end{aligned} \tag{29}$$

First we investigate the zeros of the coefficient of s^{n+p} in (29). [This coefficient corresponds to $Q_n(e^{-z})$ in Corollary 2.) For simplicity let $e^{ks} = z$. This maps the left-half s-plane into the unit circle in the z plane.

If $d_p = c_p$, then the coefficient of s^{n+p} has zeros whenever $b_n c_p = a_n c_p z^{-1}$. If a_n were zero, the coefficient in question would become constant, which satisfies the conditions of Corollary 2. If a_n is not zero, then the condition under discussion simplifies to $z^{-1} = b_n/a_n$. All solutions of this will satisfy $|z| < 1$ if $|a_n| < |b_n|$. Hence all zeros of the coefficient of s^{n+p} in (29) will lie in the left-half plane if $|a_n| < |b_n|$. This is intuitively appealing since this requires that $G(s)$ have less than unity gain at large frequencies (as required by Assumption A2).

On the other hand, if $d_p \neq c_p$, then the zeros of interest are solutions of

$$(z^{-1})^2 - \frac{2a_n d_p}{b_n(d_p - c_p)}(z^{-1}) + \frac{d_p + c_p}{d_p - c_p} = 0. \tag{30}$$

It is well-known[6] that solutions of (30) (for z^{-1}) have magnitude less than unity if and only if the following three conditions are met.

$$\left| \frac{d_p + c_p}{d_p - c_p} \right| < 1 \tag{31a}$$

$$1 + \frac{d_p + c_p}{d_p - c_p} = \frac{2d_p}{d_p - c_p} > \frac{2a_n d_p}{b_n(d_p - c_p)} \tag{31b}$$

$$\frac{2d_p}{d_p - c_p} > -\frac{2a_n d_p}{b_n(d_p - c_p)}. \tag{31c}$$

Thus conditions (31) are NAS for $| z | > 1$. Condition (31a) requires that d_p and c_p have opposite sign. Since this corresponds to terminating the line in a negative conductance at high frequencies [i.e., $\lim_{s\to\infty} Y(s) < 0$], we reject this case. If both d_p and c_p are nonzero, and have the same sign, (31a) is violated. If $d_p \neq 0$, (31b) and (31c) together require $| a_n | < | b_n |$ as before. The remaining possibility is that $d_p = 0$. This is a reasonable physical assumption; in fact, it is required by Assumption A3. Invoking Assumptions A2 and A3, the coefficient in question now becomes $b_n c_p(1 - e^{-2ks})$ which has an infinity of purely imaginary zeros, and this example no longer satisfies Corollary 2.

To see if it satisfies Corollary 1, rewrite (29) as

$$\begin{aligned} F(s) = {} & s^{n+p-1}\{(sb_nc_p + b_{n-1}c_p + b_nc_{p-1})(1 - e^{-2ks}) - 2a_nd_{p-1}e^{-ks} \\ & + b_nd_{p-1}(1 + e^{-2ks})\} \\ & + \sum_{m=0}^{n+p-2} s^m\{\sum_{i+i=m} [b_i(d_i + c_i) - 2e^{-ks}a_id_i + e^{-2ks}b_i(d_i - c_i)]\}. \end{aligned} \tag{32}$$

We complete this example by finding conditions under which the coefficient of s^{n+p-1} in (32) satisfies the conditions of Corollary 1. This coefficient can be written as

$$\begin{aligned} & ks\frac{b_nc_p}{k} + b_{n-1}c_p + b_nc_{p-1} + b_nd_{p-1} - 2a_nd_{p-1}e^{-ks} \\ & \qquad - e^{-2ks}\left(ks\frac{b_nc_p}{k} + b_{n-1}c_p + b_nc_{p-1} - b_nd_{p-1}\right). \end{aligned} \tag{33}$$

Let $w = ks$. Then (33) becomes

$$\begin{aligned} & \frac{b_nc_p}{k}\left[w + \frac{kb_{n-1}}{b_n} + \frac{kc_{p-1}}{c_p} + \frac{kd_{p-1}}{c_p}\right] \\ & \qquad - 2a_nd_{p-1}e^{-w} - \frac{b_nc_p}{k}\left[w + \frac{kb_{n-1}}{b_n} + \frac{kc_{p-1}}{c_p} - \frac{kd_{p-1}}{c_p}\right]e^{-2w}. \end{aligned} \tag{34}$$

Zeros of (34) are given by the solutions of (35)

$$(w + \alpha) + \gamma_e^{-w} - (w + \beta)e^{-2w} = 0 \tag{35}$$

where

$$\alpha = k\left(\frac{b_{n-1}}{b_n} + \frac{c_{p-1}}{c_p} + \frac{d_{p-1}}{c_p}\right)$$

$$\beta = k\left(\frac{b_{n-1}}{b_n} + \frac{c_{p-1}}{c_p} - \frac{d_{p-1}}{c_p}\right)$$

$$\gamma = -2k\,\frac{a_n}{b_n}\,\frac{d_{p-1}}{c_p}.$$

We assume that $\alpha > \beta$ since $\alpha - \beta = 2d_{p-1}/c_p k$ which is positive by Assumption A3.

It is also reasonable to assume $\alpha > 0$, since b_{n-1}/b_n must exceed zero for the denominator of $G(s)$ to be strictly Hurwitz,* and since c_{p-1}/c_p less than zero would imply zeros of $Y(s)$ in the right-half plane. These considerations also imply that $|\alpha| > |\beta|$. Using these assumptions (i.e., $\alpha > \beta$, $\alpha > 0$, $|\alpha| > |\beta|$) it follows that

$$|w + \alpha| > |w + \beta|$$

for Re $(w) \geqq 0$. Evaluating the magnitude of (35) on Re $(w) \geqq 0$ yields

$$\begin{aligned} |w + \alpha + \gamma e^{-w} - (w + \beta)e^{-2w}| &\geqq |w + \alpha| - |\gamma|\,|e^{-w}| - |w + \beta|\,|e^{-2w}| \\ &> |w + \alpha| - |\gamma| - |w + \beta|. \end{aligned}$$

If $\lim_{s\to\infty} G(s) = 0$, then $a_n = 0$ and $\gamma = 0$. This means that (35), and hence (33) and (34), have no zeros in the right-half plane and thus Corollary 1 is applicable to this problem.

VI. CONCLUSIONS

It has been shown that the time-honored technique of determining the existence of unstable poles of a closed-loop gain by counting encirclements of the critical point of the open-loop gain along a finite segment of the imaginary axis remains valid for a large class of distributed parameter systems of practical importance for which the open-loop gain approaches zero for large frequencies. Existing limitations of the Michai-

* A well-known necessary condition for a polynomial to be Hurwitz is that all coefficients have the same sign (see, for example, Ref. 6, p. 281).

lov Criterion have been removed so as to include physical systems of lossless transmission lines.

VII. ACKNOWLEDGMENTS

The author is grateful to L. D. Nelson for his careful reading of this paper and for the many discussions we had, which resulted in correcting many of the errors contained in the original manuscript.

APPENDIX

A.1 *Pontryagin's Criterion*

Let $h(z, t)$ be a polynomial with complex coefficients in the two complex variables z and t. Pontryagin[1] has developed necessary and sufficient conditions that the function $H(z) = h(z, e^z)$ have zeros with only negative real parts. We now present one of Pontryagin's main results. Let r and s be the degrees of the polynomial $h(z, t)$ with respect to z and t. Then the principal term of $h(z, t)$ is the term containing the product $z^r t^s$. Pontryagin showed that if $h(z, t)$ does not contain the principal term, then $H(z)$ has an infinity of zeros with arbitrarily large positive real parts.

Let $p(\cdot)$ and $q(\cdot)$ be real-valued functions of a real variable. We say that the zeros of these two functions alternate if: (*i*) they have no common zeros, (*ii*) they have only simple zeros, and (*iii*) between every two zeros of one of these functions there exists at least one zero of the other. The result of Ref. 1 which will be used in the present study is:

Let $h(z, t)$ be a polynomial with the principal term and $H(iy) = F(y) + iG(y)$ where $F(y)$ and $G(y)$ take on real values whenever y is real. If all zeros of the function $H(z)$ have negative real parts, then all zeros of $F(y)$ and $G(y)$ are real, alternate, and

$$G'(y)F(y) - F'(y)G(y) > 0$$

where superscript prime denotes the derivative. In order that all zeros of $H(z)$ have negative real parts, it is sufficient that all zeros of $F(y)$ and $G(y)$ are real and alternate and that $G'(y)F(y) - F'(y)G(y)$ be positive for some y.

A.2 *Proof of Theorem 1*

Choose a real number $R' > 1$ such that Re $(z) > 0$ and $F_k(z) = 0$ implies $| z | < R'$. Define the set θ,

$$\theta = \{z \mid \text{Re } (z) > 0 \quad \text{and} \quad | z | > R'.\}.$$

Let

$$D_k = \inf_{z \in \theta} | F_k(z) | > 0$$

and

$$M_k = \sup_{i=0,\cdots,k-1} \sum_{i=1}^{m} | a_{ij} |.$$

If $M_k = 0$, then $F(z) = F_k(z)$ and the theorem is trivially true. If $M_k > 0$, then for all $z \in \theta$ the following inequalities hold.

$$\begin{aligned}
| F(z) | &\geqq | z^k F_k(z) | - \sum_{j=0}^{k-1} | z^j Q_j(e^{-s}) | \\
&\geqq | z^k | D_k - \sum_{j=0}^{k-1} | z^j | \sum_{i=1}^{m} | a_{ij} | \, | e^{-\tau_{ij} s} | \\
&\geqq D_k | z^k | - \sum_{j=0}^{k-1} | z^j | \sum_{i=1}^{m} | a_{ij} | \\
&\geqq D_k | z |^k - M_k \sum_{j=0}^{k-1} | z^j | \\
&\geqq D_k | z |^k - M_k \frac{| z |^k - 1}{| z | - 1} \\
&\geqq \frac{| z |^k [D_k(| z | - 1) - M_k] + M_k}{| z | - 1} \\
&\geqq \frac{| z |^k}{| z | - 1} \left\{ [D_k | z | - (D_k + M_k)] + \frac{M_k}{| z |^k} \right\}
\end{aligned}$$

which is positive for $| z | > 1 + M_k/D_k = R_k$. Hence the magnitude of all zeros of $F(z)$ must be bounded by R, the larger of the numbers R' and R_k. The theorem follows at once by noting that $F(z)$ is analytic and that an analytic function has at most a finite number of zeros in any finite region.

A.3 *Proof of Theorem 2*

We now wish to derive an expression for the number of rhp zeros of $F(z)$. We choose a contour Γ varying along the imaginary axis from $-y$ to y (call this portion ω) where $y \geqq R_k$ of Theorem 1 and close it by a contour C outside the semicircle of radius R of Theorem 1. Let $F(z)$ and $\psi(z)$ be as defined in (15) and (16), respectively. We choose contour C (and increase y, if necessary) so that $| \psi(z) | < 1$ along C.

Let N be the number of zeros of $F(z)$ inside Γ and let

$$\Delta_\Gamma\ (F(z))$$

be the net change in arg $(F(z))$ along Γ. Then N, the number of zeros of $F(z)$ enclosed by Γ (assuming counterclockwise travel), is given by

$$N = \frac{1}{2\pi}\,\Delta_\Gamma(F(z)) = \frac{1}{2\pi}\,\Delta_\omega(F(z)) + \frac{1}{2\pi}\,\Delta_C(F(z))$$

$$= \frac{1}{2\pi}\,\Delta_\omega(F(z)) + \Delta_C(z^k) + \Delta_c(F_k(z)) + \Delta_C(1 + \psi(z))$$

since $F(z)$ has no zeros on Γ. Since $1 + \psi(z)$ does not wind around the origin (its real part being always positive), $\Delta_C(1 + \psi(z)) = 0$. Since $\Delta_C(z^k) = k/2$ we have thus proven Theorem 2.

A.4 *Sufficient conditions that Q_n does not wind around the origin*

Let

$$r_k = \min_{j=2,\cdots,m} r_j\ .$$

Then

$$\frac{e^{r_k x}}{a_{1n}}\ \mathrm{Re}\ [Q_n] = e^{r_k x} + \sum_{j=2}^{m} \frac{a_{jn}}{a_{1n}}\, e^{-(r_j - r_k)x} \cdot \cos r_j y$$

$$\frac{e^{r_k x}}{|\ a_{1n}\ |}\ |\ \mathrm{Re}\ [Q_n]\ | \geqq e^{r_k x} - \sum_{j=2}^{m} \left|\frac{a_{jn}}{a_{1n}}\right|.$$

Hence $|\ \mathrm{Re}\ [Q_n]\ | > 0$ for $x > 1/r_k \ln \alpha$ where

$$\alpha = \sum_{j=2}^{m} \left|\frac{a_{jn}}{a_{1n}}\right|.$$

Thus we need only consider a rectangle defined by $0 < x \leqq \ln \alpha/r_k$, $0 \leqq y < P$. [Note that any contour C will work if $\alpha < 1$ which is the case if assumption (11) is used.] $Q_n(e^{-z})$ does not wind around the origin for any contour with $x > \ln \alpha/r_k$, since Re $[Q_n]$ does not change sign. If $\sum_{j=1}^{m} a_{jn}^{-r_j z} \neq 0$ for $0 < x < \ln \alpha/r_k$ then Re $[Q_n]$ does not change sign for y any integer multiple of P, and $0 < x < \ln \alpha/r_k$. A suitable contour could then consist of a horizontal line at $y = KP$ from $x = 0$ to $x = \ln \alpha/r_k$, where K is an integer large enough so that $KP > R$ of Theorem 1. The rest of the contour in the first quadrant could be semicircular. The contour is completed in the fourth quadrant by the mirror image of the first quadrant. Since Re $[Q_n] \neq 0$ along this contour, $Q_n[e^{-z}]$ does not

wind around the origin. (Note that $a_{jn} > 0$, $j = 1, \cdots, m$ is sufficient to satisfy the conditions of this paragraph.) Furthermore, since Re $[Q_n]$ is even in y, $\Delta_C(Q_n(e^{-z})) = 0$ along the contour chosen.

This result can be extended to include the case where Re $[Q_n]$ has simple zeros on $y = KP$. In this case, use semicircular indentations around such points, in the direction to have Im $[Q_n] > 0$, *in both first and fourth quadrants.* (Hence the contour ceases to be symmetrical about the real axis.) Thus, along the deformed horizontal lines, the graph of $Q_n(e^{-z})$ remains in the upper-half plane. Along the semicircular portion it remains in either the right- or left-half plane. Hence no encirclements of the origin are possible and again $\Delta_C[Q_n(e^{-z})] = 0$.

REFERENCES

1. Pontryagin, L. S., "On the Zeros of Some Elementary Transcendental Functions," Am. Math. Soc. Transl., (2), 1, (1955).
2. Krall, A. M., "On the Michailov Criterion for Exponential Polynomials," SIAM Review, *8*, No. 2, (1966), pp. 184–187.
3. Reis, G. C., "On the Application of Pontryagin's Criterion to Distributed Feedback Systems," Trans. IEEE PGAC, *AC-16*, No. 2, pp. 203–204.
4. Reis, G. C., "Stability Analysis of Lumped-Distributed Feedback Systems Via Open-Loop Nyquist Plots," Sept. 1970, unpublished work.
5. Reis, G. C., "On Natural Frequencies of a Terminated Transmission Line With Feedback," Trans. IEEE PGCT, *CT-18*, No. 5, pp. 569–570.
6. Cheng, D. K., *Analysis of Linear Systems*, Reading, Mass.: Addison-Wesley Publishing Co., Inc., 1959, pp. 329–336.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

The Spectral Density of a Coded Digital Signal

By B. S. BOSIK

(Manuscript received August 23, 1971)

The stochastic process appearing at the output of a digital encoder is investigated. Based upon the statistics of the code being employed, a systematic procedure is developed by means of which the average power spectral density of the process can be determined. The method is readily programmed on the digital computer, facilitating the calculation of the spectral densities for large numbers of codes. As an example of its use, the procedure is applied in the case of a specific multi-alphabet, multi-level code.

I. INTRODUCTION

In recent years, increased interest has been focused on more complex multi-alphabet, multi-level codes.[1-4] Such codes are designed to produce a digital pulse train with specific spectral properties making it suitable for transmission over digital repeatered lines. These properties generally include the absence of a dc component and a strong spectral component from which timing can be extracted. This paper presents a method for calculating the spectral composition of the pulse trains resulting from the use of these codes. The procedure is applicable to a wide variety of codes.

A code may be defined as a set of mappings from a set of input symbols (or words) to a set of codewords. Each mapping is called an alphabet. The code may use different alphabets depending upon the state of the coded signal.[1] It is desirable for unique decipherability that the set of mappings be one-to-one, i.e., that no matter to how many alphabets a codeword belongs, it always corresponds to the same input symbol. However, this restriction will not be imposed here.

In general, when the code is applied to a sequence of input symbols, the resulting encoded signal is a stochastic process, the statistics of which depend on the input symbol sequence statistics and the code statistics. For convenience, a random input symbol sequence will be assumed so that the input symbols are equally likely. Even if the

symbols are not equiprobable, the procedure for calculating the spectral density outlined in Section III remains valid, although the methods for calculating the required signal statistics described in Section IV must be modified. The spectral density derivation only becomes inapplicable when the statistics of the input symbols vary with time, so that the ergodicity and stationarity assumptions of Sections II and III are no longer valid.

The codes to be considered in this paper will have N states, each state corresponding to a single alphabet. The alphabet assignment need not be unique, i.e., more than one state can correspond to the same alphabet. The codes will have a block length L and the number of codeword symbol values (levels) will be M.

II. THE CODED SIGNAL

The codeword symbols are, in general, transmitted on some standard pulse shape $g(t)$ at intervals of duration T. The signal, then, may be expressed by

$$x(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT) \tag{1}$$

where a_n is the codeword symbol value for the time slot $nT \leqq t \leqq (n + 1)T$. The values which $\{a_n\}$ assume are determined by the code and the input symbols which are to be coded. The discrete parameter random process formed by the sequence of codeword symbols $\{a_n\}$ has an autocorrelation function $R(k) = E\{a_n a_{n+k}\}$, and, thus, is assumed to be wide-sense stationary. The autocorrelation function of the coded signal $x(t)$ is, then,

$$R_x(t + \tau, t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} R(m - n) g(t + \tau - nT) g(t - mT) \tag{2}$$

which is, in general, a function of both t and τ. The coded signal is not, therefore, wide-sense stationary. However, it is easily shown that $R_x(t + \tau, t)$ is periodic in t with period T. The coded signal is, then, a cyclostationary process.

III. THE POWER SPECTRAL DENSITY OF THE SIGNAL

Since the coded signal is not a wide-sense stationary random process, the Fourier transform relationship between the autocorrelation function and the power spectral density cannot be invoked to find its power spectrum. However, the average power spectral density of a cyclo-

stationary process of the form described above has been derived by W. R. Bennett.[5] Under the assumption that the process is ergodic, the spectral density is

$$w_x(f) = \frac{1}{T} \mid G(f) \mid^2 \left[R(0) + 2 \sum_{k=1}^{\infty} R(k) \cos 2\pi k f T \right] \tag{3}$$

where $G(f)$ is the Fourier transform of $g(t)$. The determination of the spectral density, then, requires the calculation of the autocorrelation function $R(k)$.

A method of calculating $R(k)$ can be derived as follows. Let the probability of being in state i during time slot n be

$$P(s_n = S_i) = P(S_i); \quad i = 1, 2, \cdots, N \tag{4}$$

and let the probability that the symbol a_n assumes the value A_l, $l = 1, 2, \cdots, M$, and a_{n+k} assumes the value A_m, $m = 1, 2, \cdots, M$, given that $s_n = S_i$ be

$$P(a_n = A_l, a_{n+k} = A_m \mid s_n = S_i) = P_k(A_l, A_m \mid S_i). \tag{5}$$

Then $R(k)$ can be expressed as

$$R(k) = \sum_{i=1}^{N} P(S_i) \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m P_k(A_l, A_m \mid S_i). \tag{6}$$

But in time slot n, there is a probability of $1/L$ of being in the jth symbol, a_j, of a codeword, $j = 1, 2, \cdots, L$. Thus,

$$P_k(A_l, A_m \mid S_i) = \sum_{j=1}^{L} \frac{1}{L} P_k(A_l, A_m \mid S_i, j). \tag{7}$$

Substituting eq. (7) in eq. (6), we obtain

$$R(k) = \frac{1}{L} \sum_{i=1}^{N} P(S_i) \sum_{j=1}^{L} \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m P_k(A_l, A_m \mid S_i, j). \tag{8}$$

Now define

$$\begin{aligned} R_{S_i}(j, j+k) &= \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m P_k(A_l, A_m \mid S_i, j) \\ &= E_{S_i}\{a_j a_{j+k}\}; \qquad j = 1, 2, \cdots, L. \end{aligned} \tag{9}$$

Thus,

$$R(k) = \frac{1}{L} \sum_{i=1}^{N} P(S_i) \sum_{j=1}^{L} R_{S_i}(j, j+k). \tag{10}$$

The equation by which $R_{S_i}(j, j+k)$ is calculated is a function of

both j and k. The values of j and k determine the relative positions of the codewords to which the symbols a_i and a_{i+k} belong. This knowledge combined with a knowledge of the coded signal statistics allows the calculation of $R_{S_i}(j, j + k)$ as follows:

(*i*) $k = 0$:

For $k = 0$, a_j and a_{i+k} are in the same position of the same codeword. Thus,

$$R_{S_i}(j, j) = \sum_{l=1}^{M} A_l^2 P(A_l | \ S_i \ , j) \tag{11}$$

where

$$P(A_l \mid S_i \ , j) = P(a_j = A_l \mid s_j = S_i). \tag{12}$$

(*ii*) $k = 1, 2, \cdots, L$:

For this range of k, a_{i+k} is in the same codeword as a_i for $j \leqq L - k$, and is in the next codeword for $j > L - k$. Thus,

$$R_{S_i}(j, j + k) = \begin{cases} \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m P_k(A_l \ , A_m \mid S_i \ , j), & j \leqq L - k \\ \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m \sum_{n=1}^{N} P(A_l \ , S_n \mid S_i \ , j) & \\ \qquad \cdot P(A_m \mid S_n \ , j + k - L), & j > L - k \end{cases} \tag{13}$$

where

$$P(A_l \ , S_n \mid S_i \ , j) = P(a_j = A_l \ , s_{j+L} = S_n \mid s_j = S_i). \tag{14}$$

(*iii*) $k = L + 1, L + 2, \cdots, 2L$:

For this range of k, a_{i+k} is in the codeword immediately following the codeword containing a_i for $j \leqq 2L - k$, and is two codewords away for $j > 2L - k$. Thus,

$$R_{S_i}(j, j + k) = \begin{cases} \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m \sum_{n=1}^{N} P(A_l \ , S_n \mid S_i \ , j) & \\ \qquad \cdot P(A_m \mid S_n \ , j + k - L), & j \leqq 2L - k \\ \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m \sum_{n=1}^{N} P(A_l \ , S_n \mid S_i \ , j) \sum_{p=1}^{N} P_1(S_p \mid S_n) & \\ \qquad \cdot P(A_m \mid S_p \ , j + k - 2L), & j > 2L - k \end{cases} \tag{15}$$

where

$$P_1(S_p \mid S_n) = P(s_{i+L} = S_p \mid s_i = S_n). \tag{16}$$

(*iv*) $k = QL + 1, QL + 2, \cdots, (Q + 1)L$ for $Q \geqq 2$:

For this range of k, a_{i+k} is in Qth codeword following the codeword containing a_i for $j \leqq QL - k$, and is $Q + 1$ codewords away for $j > QL - k$. Thus,

$$R_{S_i}(j, j + k) = \begin{cases} \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m \sum_{n=1}^{N} P(A_l, S_n \mid S_i, j) \sum_{p=1}^{N} P_{Q-1}(S_p \mid S_n) \\ \qquad \cdot P(A_m \mid S_p, j + k - QL), \qquad j \leqq QL - k \\ \sum_{l=1}^{M} \sum_{m=1}^{M} A_l A_m \sum_{n=1}^{N} P(A_l, S_n \mid S_i, j) \sum_{p=1}^{N} P_Q(S_p \mid S_n) \\ \qquad \cdot (A_m \mid S_p, j + k - (Q + 1)L), \qquad j > QL - k \end{cases} \tag{17}$$

where

$$P_Q(S_p \mid S_n) = \sum_{i=1}^{N} P_{Q-1}(S_i \mid S_n) P_1(S_p \mid S_i), \qquad Q \geqq 2 \tag{18}$$

and can be calculated recursively.

In summary, then, the procedure described in eqs. (11) through (18) is used to calculate $\{R_{S_i}(j, j + k)\}$ which is substituted in eq. (10) to obtain $\{R(k)\}$. The spectral density $w_x(f)$ can then be obtained from $\{R(k)\}$ by means of eq. (3).

It is easily seen that for any code other than the most trivial, the calculation of the spectral density is a formidable task. However, it is a relatively straightforward procedure to program a digital computer to perform the above calculations; and this is the most profitable use of the procedure.

IV. THE CODED SIGNAL STATISTICS

The calculation of the spectral density described above requires the knowledge of numerous probabilities concerning the code. These statistics are, in general, readily obtainable from the code by merely counting the number of occurrences of the phenomenon involved, or are determined from simple calculations involving previously obtained probabilities. The procedure for obtaining each of the necessary probabilities follows:

(*i*) The State Transition Probabilities:

$$P_1(S_p \mid S_n) = P(s_{i+1} = S_p \mid s_i = S_n); \quad p, n = 1, 2, \cdots, N. \tag{19}$$

This transition probability is obtained by counting the number of codewords in the alphabet used when in state S_n, whose next state is S_p. (Given any codeword in any state, the next state is uniquely defined at the end of that codeword.) The resulting sum is divided by the number of codewords in the alphabet. The probability of a transition from state S_n to S_p in Q steps, $P_Q(S_p \mid S_n)$, can be calculated recursively from $P_1(S_p \mid S_n)$ via eq. (18).

(*ii*) The State Probabilities:

$$P(S_i) = P(s_n = S_i); \quad i = 1, 2, \cdots, N. \tag{20}$$

The probability of being in state S_i is calculated from $\{P_1(S_p \mid S_n)\}$. $\{P(S_i)\}$ are the solutions to the set of simultaneous linear equations

$$\sum_{k=1}^{N} P(S_k)P(S_i \mid S_k) = P(S_i); \qquad j = 1, 2, \cdots, N. \tag{21}$$

However, only $N - 1$ of these equations are linearly independent. An additional equation must be used:

$$\sum_{i=1}^{N} P(S_i) = 1. \tag{22}$$

(*iii*) The Symbol Probabilities:

$$P(A_l \mid S_i, j) = P(a_j = A_l \mid s_j = S_i); \qquad \begin{aligned} i &= 1, 2, \cdots, N \\ j &= 1, 2, \cdots, L \\ l &= 1, 2, \cdots, M. \end{aligned} \tag{23}$$

These probabilities are determined by counting the number of occurrences of symbol A_l in the jth position of the codewords in the alphabet used when in state S_i. This sum is divided by the number of codewords in the alphabet.

(*iv*) The Symbol Combination Probabilities:

$$\begin{aligned} &P_k(A_l, A_m \mid S_i, j) \\ &\quad = P(a_j = A_l, a_{j+k} = A_m \mid s_n = S_i); \qquad \begin{aligned} i &= 1, 2, \cdots, N \\ j &= 1, 2, \cdots, L \\ l, m &= 1, 2, \cdots, M. \\ k &\leqq L - j \end{aligned} \end{aligned} \tag{24}$$

This is the probability of the occurrence of the symbols A_l and A_m, in time slots j and $j + k$ respectively, within the codewords of the alphabet used in state S_i (i.e., $k \leqq L - j$). Thus, for position j of the alphabet of S_i, the number of times $a_j = A_l$ and $a_{j+k} = A_m$ in the same codeword are counted, and divided by the number of codewords in the alphabet.

(v) The Conditional State Transition Probabilities:

$$P(A_l, S_n \mid S_i, j) = P(a_j = A_l, s_{j+L} = S_n \mid s_j = S_i); \quad i, n = 1, 2, \cdots, N \quad j = 1, 2, \cdots, L \quad l = 1, 2, \cdots, M. \quad (25)$$

This is the probability that a codeword in the alphabet of state S_i has the symbol A_l in the jth position and has the state S_n as its next state. Thus, the number of times that a codeword is in the alphabet of state S_i, whose next state is S_n and whose symbol level is A_l in the jth position, are counted, and divided by the number of codewords in the alphabet.

These five sets of statistics are all that are required to perform the calculation of the spectral density. Although following the procedures for obtaining these probabilities is a very straightforward task, it is, again, a tedious one, especially for any reasonably complex code. Here again, the digital computer can be used to good advantage.

V. AN EXAMPLE—THE FRANASZEK MS-43 CODE

The Franaszek MS-43 code[1] is a ternary, 4-state, 3-alphabet code of word length 3. It is one of a family of codes designed to produce a digital pulse train with specific desirable properties. These properties include the absence of a dc component, a bounded sum of previous digits, and a strong spectral component from which the signaling frequency can be derived. The code is shown in Table I. Alphabet R_1 is used when in state S_1, alphabet R_2 is used in state S_2 or S_3, and alphabet R_3 is used in state S_4. The state is determined at the end of a codeword by summing all previous digital symbols. This sum is inherently restricted to be 1, 2, 3, or 4 corresponding to states S_1, S_2, S_3, and S_4. Tables II through VI list the statistics necessary for calculating the spectral density as determined by the procedures described in Section IV. The digital computer was utilized to perform the spectral

TABLE I—THE MS-43 CODE

Binary Input Words	R_1	R_2	R_3
0000	+++	−+−	−+−
0001	++0	00−	00−
0010	+0+	0−0	0−0
0100	0++	−00	−00
1000	+−+	+−+	−−−
0011	0−+	0−+	0−+
0101	−0+	−+0	−0+
1001	00+	00+	−−0
1010	0+0	0+0	−0−
1100	+00	+00	0−−
0110	−+0	−+0	−+0
1110	+−0	+−0	+−0
1101	+0−	+0−	+0−
1011	0+−	0+−	0+−
0111	−++	−++	−−+
1111	++−	+−−	+−−

TABLE II—STATE TRANSITION PROBABILITIES FOR MS-43 CODE

$P_1(S_p|S_n)$

	P			
n	1	2	3	4
1	6/16	6/16	3/16	1/16
2	5/16	6/16	5/16	0
3	0	5/16	6/16	5/16
4	1/16	3/16	6/16	6/16

TABLE III—STATE PROBABILITIES FOR MS-43 CODE

$P(S_i)$

i	$P(S_i)$
1	5/28
2	9/28
3	9/28
4	5/28

density calculation. The resulting normalized spectrum is plotted in Fig. 1. The result is consistent with the expected properties of the coded signal spectrum, i.e., zero dc component and periodicity with period $1/T$.

VI. CONCLUSION

A general procedure for determining the average power spectral density of a coded digital signal has been presented. The procedure is long, but straightforward and readily programmable on the digital computer. With the aid of the computer, the spectral content of signals resulting from the implementation of large numbers of codes can be obtained.

VII. ACKNOWLEDGMENTS

The author is indebted to W. R. Bennett whose advice and encouragement made this work possible. The author also had the benefit

TABLE IV—SYMBOL PROBABILITIES FOR MS-43 CODE

$P(A_m|S_i, j) = P(A_m|S_i), \forall j = 1, 2, 3$

	S_i			
A_m	S_1	S_2	S_3	S_4
+	8/16	5/16	5/16	3/16
−	3/16	5/16	5/16	8/16

TABLE V—SYMBOL COMBINATION PROBABILITIES FOR MS-43 CODE

$P_k(A_\ell, A_m|S_i, j)$

$j = 1, k = 1$

	S_i			
$A_\ell A_m$	S_1	S_2	S_3	S_4
++	3/16	0	0	0
+−	2/16	3/16	3/16	2/16
−+	2/16	3/16	3/16	2/16
−−	0	0	0	3/16

$j = 1, k = 2$

	S_i			
$A_\ell A_m$	S_1	S_2	S_3	S_4
++	3/16	1/16	1/16	0
+−	2/16	2/16	2/16	2/16
−+	2/16	2/16	2/16	2/16
−−	0	1/16	1/16	3/16

$j = 2, k = 1$

	S_i			
$A_\ell A_m$	S_1	S_2	S_3	S_4
++	3/16	3/16	1/16	0
+−	2/16	2/16	2/16	2/16
−+	2/16	2/16	2/16	2/16
−−	0	1/16	1/16	3/16

Table VI—Conditional State Transition Probabilities for MS-43 Code

$P(A_\ell, S_n|S_i, j)$

		S_i											
		S_1			S_2			S_3			S_4		
		j											
A_ℓ	S_n	1	2	3	1	2	3	1	2	3	1	2	3
+	S_1	2/16	2/16	2/16	1/16	1/16	0	0	0	0	0	0	0
+	S_2	3/16	3/16	3/16	2/16	2/16	2/16	1/16	1/16	0	0	0	0
+	S_3	2/16	2/16	2/16	2/16	2/16	3/16	2/16	2/16	2/16	1/16	1/16	1/16
+	S_4	1/16	1/16	1/16	0	0	0	2/16	2/16	3/16	2/16	2/16	2/16
−	S_1	2/16	2/16	2/16	2/16	2/16	3/16	0	0	0	1/16	1/16	1/16
−	S_2	1/16	1/16	1/16	2/16	2/16	2/16	2/16	2/16	3/16	2/16	2/16	2/16
−	S_3	0	0	0	1/16	1/16	0	2/16	2/16	2/16	3/16	3/16	3/16
−	S_4	0	0	0	0	0	0	1/16	1/16	0	2/16	2/16	2/16

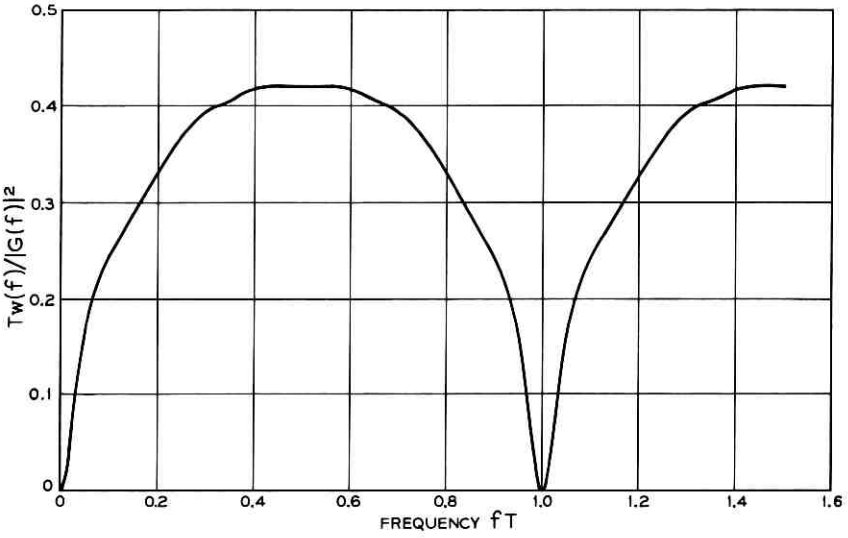

Fig. 1—Power spectral density for the Franaszek MS-43 code.

of a private communication from A. Fromageot discussing the problem of calculating the spectral density of the Franaszek MS-43 code.

REFERENCES

1. Franaszek, P. A., "Sequence-State Coding for Digital Transmission," B.S.T.J., *47*, No. 1 (January 1968), pp. 143–157.
2. Sipress, J. M., "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," IEEE Trans. Commun. Tech, *COM-13*, (September 1965), pp. 366–372.
3. Neu, W., and Kündig, A., "Project for a Digital Telephone Network," IEEE Trans. Commun. Tech, *COM-16*, (October 1968), pp. 633–648.
4. Johannes, V. I., Kaim, A. G., and Walzman, T., "Bipolar Pulse Transmission with Zero Extraction," IEEE Trans. Commun. Tech., *COM-17*, No. 2 (April 1969), pp. 303–310.
5. Bennett, W. R., "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Objective Measures of Peak Clipping and Threshold Crossings in Continuous Speech

By PAUL T. BRADY

(Manuscript received September 1, 1971)

This study reports data on the statistics of instantaneous speech levels in continuous speech samples, with special emphasis on threshold crossings and other quantities related to peak clipping. All clipping thresholds for each speech sample are defined with respect to the individual speech level for that sample, specified in equivalent peak level (epl). Speech clipping is also treated as speech-correlated noise by assuming that it is caused not by a voltage limiting process, as actually occurs, but by an additive "phantom signal" that will cause the original signal to appear to be clipped. Empirical measures are obtained for the percent time a clipping level is exceeded, for the relation between phantom signal (i.e., noise) power and clipping level, and for the loss in signal power resulting from clipping. The relation is also established between epl and average (rms) power to allow signal power and signal-to-clipping noise levels to be specified.

I. INTRODUCTION

This study reports data on the statistics of instantaneous speech levels in continuous speech samples, with special emphasis on threshold crossings and other quantities related to peak clipping. Since a standard measure of specifying the amount of peak clipping is lacking, the study begins by arbitrarily defining the clipping level in terms of an objective speech level measure, the equivalent peak level (epl).[1] Then, measures of time spent above various thresholds (or clipping levels) are reported. Finally, a method is suggested for interpreting clipping as speech-correlated noise. Empirical measures are given for the amount of noise introduced as a function of the amount of clipping, and signal power lost because of clipping.

This paper does not address itself to subjective measurements related to clipping of the speech signal. The subjective effects of clipping must

eventually be considered in a treatment of the broad problem of determining clipping performance objectives. It is also necessary to examine the objective effects of clipping on the transmitted signal so that guidelines can be established for design and operation of transmission systems. This paper is directed toward providing data to assist in engineering considerations of speech circuits.

II. DEFINITION OF A MEASURE OF PEAK CLIPPING

Peak clipping, produced by an abrupt limiting of a waveform when its amplitude attempts to exceed a clipping voltage, is possible in virtually any speech transmission system if the speech level becomes too high. It is quite easy to specify the limiting voltage in an absolute sense (e.g., ± 0.5 V), but the speech impairments caused by clipping are a function of the relation between the clipping voltage and the speech level.

One clipping measure sometimes used is the difference between the clipping level and the volume unit (VU). This difference is quite variable because of the variability in the VU measure, as determined in experiments by Shearme and Richards,[2] and Brady.[3] A study related to VU and speech peaks was done by Noll,[4] who measured the difference between the highest instantaneous peak of a speech sample and the VU for possible application to peak clipping. This difference was also quite variable among samples, causing Noll to conclude that "volume (i.e., VU) distributions cannot easily be converted to peak distributions." He states further that the variability "might have been caused by inaccuracies inherent in reading a VU meter."

Another clipping measure sometimes used is the difference between the highest instantaneous peak in the speech sample and the clipping voltage. Although this measure is objective and fairly easily obtained, it is too dependent on the near chance value of the highest peak. Should this peak be due to an unusually loud segment of speech or even a click or spurious signal, the clipping measure has little relationship to the major part of the speech sample.

The present study begins by defining a new measure of clipping as the difference between the epl of a speech sample and the clipping level. This difference is measured in dB. It promises to be more stable than VU-related clipping measures since the epl is an objective measure with considerably less variability than the VU.[1,3] It is emphasized that this new clipping measure is not proposed here as a new standard; it is simply adopted here as a more stable basis than previous measures so that subsequent derived measures will be more precisely defined.

There are three principal derived measures discussed here. One measure is a count of the percent of time that the clipping level is exceeded. For example, if one clips 3 dB below the epl, what percent of the waveform would be affected? The second is the measure of power in a "phantom" interference signal. The clipping is assumed to be caused not by a voltage limiting process, but rather by a second interference signal added to the first. By estimating power in the phantom signal, a signal/noise ratio can be defined for different clipping levels. The third measure is power lost due to clipping.

III. USE OF CONTINUOUS SPEECH

Many objective measures of speech are strong functions of the "activity factor" or the percent of time a person is talking. For example, a person speaking at a fixed level can change his average power by varying his activity factor. In the present study, the quantities to be measured are dependent on the time base chosen.

In studies on clipping, there is little interest in what happens when speech is not present. It makes little sense to examine long silent intervals, which would occur during a telephone conversation. Therefore, this study's measures are restricted to only those times when speech is present.

The author knows of no speech detecting technique that will define intervals of speech activity in a manner that is insensitive to arbitrary choices of parameters such as detector threshold setting. This is shown, for example, in two previous studies in which substantial variation in detected speech patterns occurred with fairly small changes in detector parameters.[5,6] In the present study, the detection process is bypassed by using "continuous speech" material that contains a negligible number of perceived gaps.

The method of recording continuous speech is documented in Ref. 6. These recordings were prepared from recordings of experimental telephone conversations by manually splicing together sections containing a continuous flow of speech. There were eight male and eight female speakers each providing two separate recordings, except for two women who each made only one recording. In all, there were 16 male and 14 female continuous speech samples, the average length of each sample being about 55 seconds.

Measures made on these samples will be considered valid measures made "during speech." For example, if a threshold is exceeded here 3 percent of the time, this will imply that in noncontinuous "real" speech, the threshold is exceeded 3 percent "during speech" or "while the

speaker is talking." The continuous speech tapes were edited by the author using a highly empirical procedure and personal judgment. Therefore, all measures reported here are dependent on the author's personal judgment, to the extent that the selection of continuous speech material is dependent on this judgment.

The 30 samples of continuous speech provide a basis for studying the variability of the desired measures over different *speech samples.* Note again that only 16 speakers are represented in the 30 samples, so that measures of variability are somewhat confounded if they are to be applied to variability over *different speakers.* This confounding is probably slight, since two samples from the same speaker are sometimes quite different in measures such as speech level and dominance of conversation (whether the person is basically listening or talking).

The speech samples of all the men were played back at a level high enough to use nearly the full range of a 12-bit A-to-D converter. Once the level adjustment for males was set, it was not changed from sample to sample. The level adjustment was then reset for the women, and all female speech samples were played at the new adjustment.

IV. RELATION OF EPL TO PEAKS

The epl reads the peak of an *assumed* log uniform distribution of instantaneous voltage, whose rms voltage above threshold is the same as that measured in the sample.* If speech were distributed exactly according to the log uniform distribution, then the epl would be the highest level or the "peak." In practice, the speech distribution does not abruptly and neatly terminate at some fixed level. Occasionally there are voltages exceeding the range of levels of the speech. The expected epl would be a "nominal" upper voltage limit, which would be exceeded occasionally during loud speech passages.

To get a quantitative measure of the relation between the epl and the highest instantaneous peaks, all 30 continuous speech samples were played into an A-to-D converter connected to a PDP-8 computer. The computer stored a histogram count of the measured voltages sampled at 1 kHz. After calculating the epl at the end of each speech sample, the number of A-to-D readings that exceeded the epl was counted.

On the average, the epl was exceeded 2.0 percent of the time (σ = 0.29 percent) for female speech samples and 2.9 percent of the time

* The epl as defined in Ref. 1 has been revised as follows. Step 5a should read, "if $D \leq 6.75$, set $\Delta = (D - 2.75)/0.4$."

(σ = 0.46 percent) for male speech samples. The average of all 30 samples was 2.5 percent. In general, in noncontinuous or "standard" speech, the epl would be exceeded about 2.5 percent of the "time that speech is present."

The speech voltages that did exceed the epl occasionally reached levels of 7 or even 8 dB above the epl, but such events were very rare. The epl + 6 dB level was exceeded about one-tenth as often as the epl. That is, the epl + 6 dB was typically exceeded between 0.2 and 0.3 percent of the time. Thus, the epl + 6 dB appears to be a practical upper limit of the instantaneous speech levels.

V. PERCENT TIME CLIPPING LEVEL IS EXCEEDED

The log-uniform speech distribution model predicts that as a threshold (i.e., clipping level) is lowered, equal dB decreases in the threshold produce equal increments in percent time that speech exceeds the threshold.[7] This model is somewhat inaccurate in the immediate region of the epl, but it has been found to hold up over a wide range (30 or 40 dB) below the epl of most speech samples tested.[1,7] Eventually, for very low thresholds, the model must break down because one could continue lowering the threshold in equal dB increments to minus infinity; and one certainly cannot indefinitely add equal percent time increments.

Two experiments with continuous speech yielded data relevant to percent time above clipping level. The first showed that the epl itself is cleared about 2.5 percent of the time. This is one point on the "percent time vs clipping level" curve.

The second experiment sought to establish another point on the curve by obtaining a scatter plot of percent time over a range of thresholds for different speech samples. In an earlier study, percent time over a fixed threshold of −25 dBm was measured for 30 continuous speech samples. However, the samples were played at their original recorded levels rather than at levels equalized for A-to-D converter range. The epls had a sigma of 3.6 dB, which was a larger range than those obtained for the samples with equalized levels. If the epls had a fairly large range, it follows that epl minus threshold also had the same large range for a fixed threshold.

The scatter plot of percent time over threshold vs threshold (re epl) is shown in Fig. 1. The mean value of the 30 points occurs at the coordinates 15 dB for epl minus threshold, and 30 percent for time above threshold. Perhaps the simplest way to provide a linear fit to

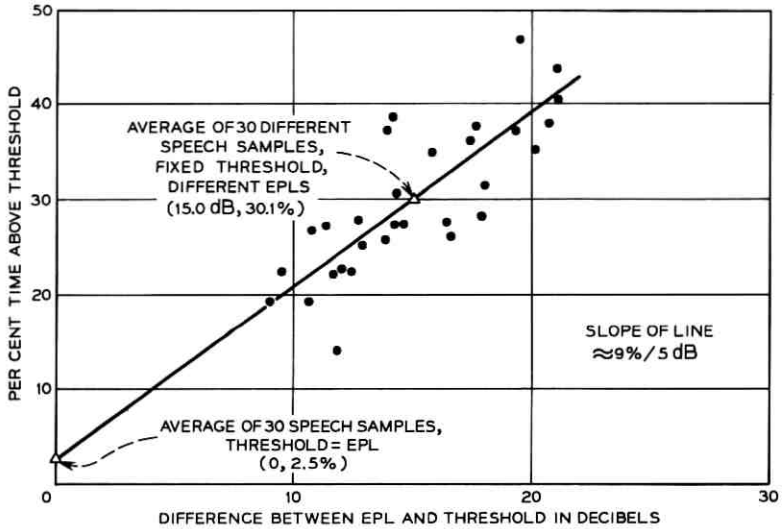

Fig. 1—Percent time threshold is cleared vs difference between epl and threshold for continuous speech.

the Fig. 1 data is to draw a straight line between the mean values for the two experiments.

Each of these two points determining the linear fit is an average of 30 samples. An alternative means of fitting a line to the data could be to use all 60 points in a least-mean-square fit. This might yield a more "accurate" fit to these data, but it must be remembered that the vertical axis, the "percent time" measure, is based upon an empirical definition of continuous speech. The straight line connecting the means of the two data sets is probably just as reasonable a fit if the curve is to be regarded as an approximate guide for application to speech circuit design. Note that in Fig. 1, the line passes within 5 percent divisions of most of the points, tending to justify the linear fit predicted by the idealized log-uniform speech distribution model. Previous work suggests that the linearity will be maintained until 30 or 35 dB below the epl.[7]

The line has a slope of 1.83 percent increase in time for every 1 dB drop in threshold, or approximately 9 percent for every 5 dB drop. This result can be combined with the conclusion of Section IV as a guideline for threshold clearance percentages:

Instantaneous voltages rarely occur at a level higher than 6 dB above the epl. During *continuous speech,* the epl is exceeded roughly 2.5 percent of the time. A threshold is cleared an additional 9 percent of the time for every 5 dB drop in threshold below the epl.

VI. RELATION OF EPL TO AVERAGE POWER

While obtaining the epls of the 30 speech samples, measures were also obtained of the long-term rms power in the continuous speech. The average epl minus rms for the 30 samples was 10.0 dB, with a standard deviation of 0.8 dB.

Therefore, during continuous speech, the speech power is about 10 dB below the epl. This result will be useful in the next section in relating speech power to clipping signal power.

VII. POWER IN A PHANTOM CLIPPING WAVEFORM

Instead of thinking of peak clipping as a limiting process, imagine it to be caused by the addition of a second phantom signal that will cause the original waveform to be restrained at some voltage. This process is illustrated in Fig. 2, showing the original signal, the phantom signal, and the result of adding them (i.e., the clipped signal).

The phantom signal can be considered as speech-correlated noise. If its power is known, then a signal/noise (S/N) ratio for clipped speech can be determined. In applying the S/N ratio, be careful to remember that the noise is speech-correlated, and hence may produce very different

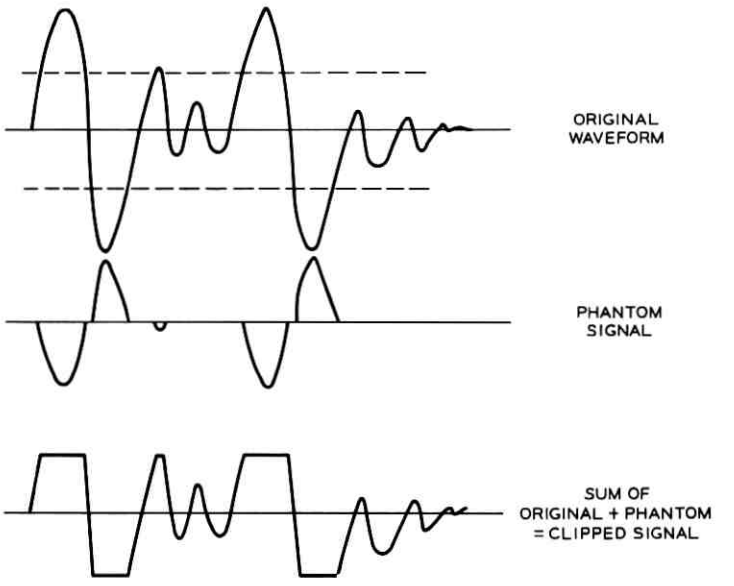

Fig. 2—Illustration showing how clipping can be thought of as due to an additive phantom signal.

effects on intelligibility, detectability, etc., than would a steady background noise with the same S/N ratio.

Note that this "power" of the phantom signal is a fictitious quantity, and no such *power* is added to the signal, even though a *voltage* is added. In fact, clipping produces a net power *loss* since it removes part of the signal.

The phantom signal power was calculated by playing a speech sample into the PDP-8 computer, and recording a histogram of the absolute magnitudes of all voltage levels obtained. When finished, the computer chose a fixed threshold level, T, and for each voltage level above this computed (Voltage − Threshold)^2. An accumulator was updated with this quantity n_T times, where n_T was the number of counts for that threshold.

The computer then divided the total $\sum (V - T)^2$ by the total number of A-to-D samples for the *entire* speech sample. The time base for averaging was the total length of the continuous speech sample, not the number of voltages exceeding T. Thus, the power for the phantom signal (the $V - T$ voltage waveform) was an average "during the time speech is present." The lower the threshold, the greater the power in the $(V - T)$ phantom noise signal would be, since $(V - T)$ was greater for each voltage and more voltages exceeded T.

The above process was done for six thresholds: -5, -10, $\cdots$, -30 dBm. For each speech sample a curve of phantom power vs clipping voltage was produced. These curves for all 30 speech samples are plotted in Fig. 3. Each speech sample curve has been normalized to its epl.

The author knows of no speech distribution model that would be appropriate for predicting the curves of Fig. 3. The log-uniform model, which works well for relating power to epl, is a poor fit in the voltage region close to the epl, since it incorrectly predicts an abrupt upper limit to the speech voltages at the epl. Lacking a model, the curves of Fig. 3 were treated as a scatter plot of independent points (each curve produced 5 or 6 such points) which were fitted with a third-degree polynomial. The curve fitting program, supplied by E. A. Youngs of Bell Laboratories, yielded the curve described by the equation below. In this equation, P = power in the phantom signal in dB (re epl) and T = clipping threshold in dB (re epl). (E.g., for clipping level 1 dB below the epl, $T = -1$.)

For all speakers,

$$P = -21.86 - 1.193T - 0.0443T^2 - 0.00057T^3.$$

This curve is plotted in Fig. 4.

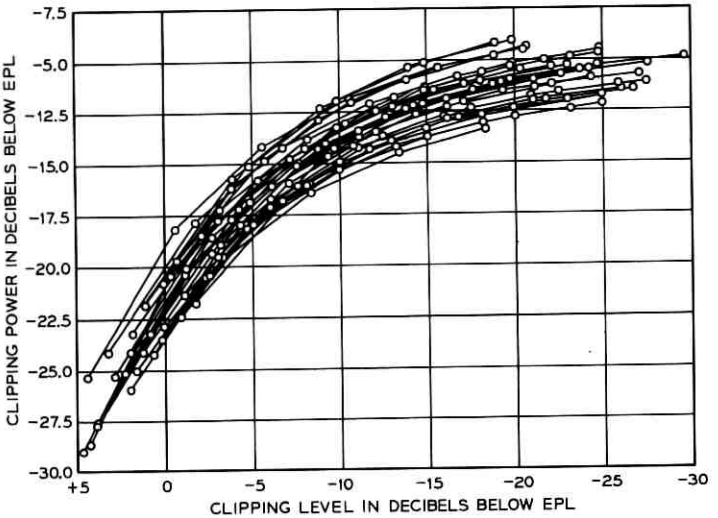

Fig. 3—Clipping power in a phantom clipping signal as a function of clipping level, for 30 continuous speech samples. Each curve is normalized to the epl for the associated speech sample.

VIII. POWER LOST FROM PEAK CLIPPING

If a signal is clipped, the resulting signal has less power than the original signal. The power loss in dB was calculated for all 30 continuous speech samples using six clipping levels for each sample. The 30 curves of power loss vs clipping level are superimposed in Fig. 5. The polynomial fit to these data, shown in Fig. 6, is given by

$$P = -1.084 + 0.373T - 0.012T^2,$$

where P = power loss (if 5 dB is lost, $P = -5$) and T = clipping threshold (if 5 dB below epl, $T = -5$). (A second-order polynomial was sufficient to meet the "fit criterion" of the polynomial curve fitting program.)

Also plotted in Fig. 6 is a curve of power loss vs clipping obtained by Wathen-Dunn and Lipke,[8] after an empirical instantaneous speech level probability distribution function developed by Davenport.[9] Wathen-Dunn defines his zero dB clipping level reference point as that point exceeded by only 0.1 percent of Davenport's speech signal, which consisted of having speakers read aloud from a book. Although such speech must have contained perceptible pauses, it is closer to "continuous speech" than a telephone conversation. To compare Wathen-Dunn's work with this, let us consider his 0.1 percent peak point for book

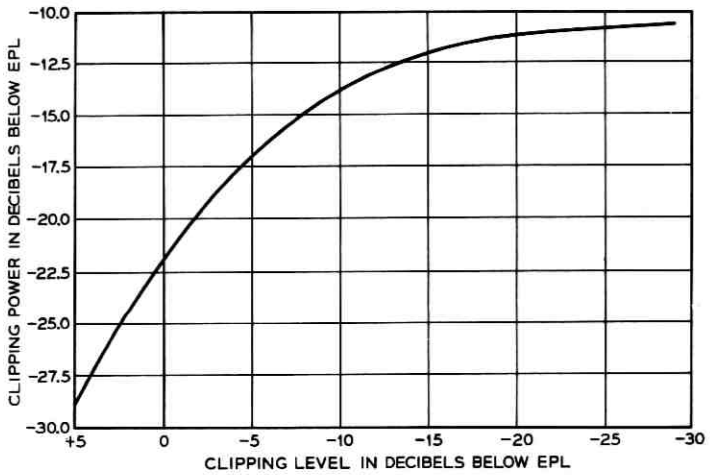

Fig. 4—Third-order polynomial fit to data of Fig. 3.

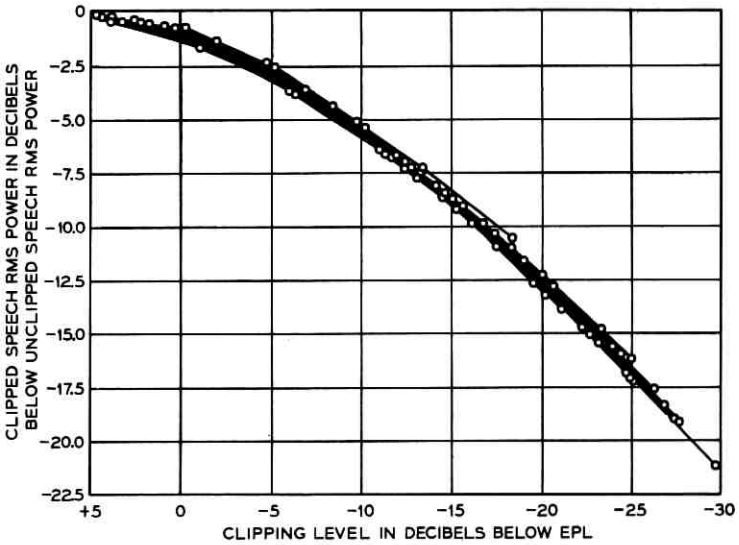

Fig. 5—Power loss vs clipping level for 30 samples of continuous speech.

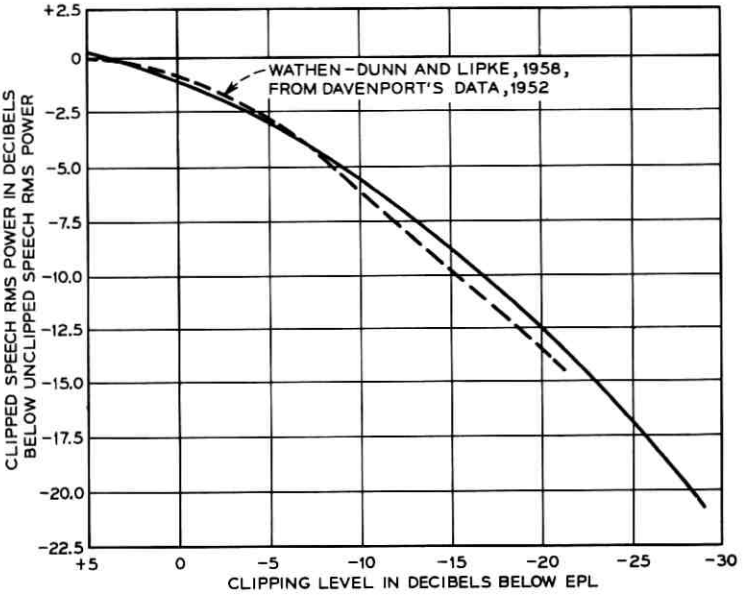

Fig. 6—Solid curve is a polynomial fit to data of Fig. 5. Dashed curve is taken from Wathen-Dunn and Lipke.[8]

reading as equivalent to the epl + 6 dB "peak" criterion established here for continuous speech. Fig. 6 of the present report is Wathen-Dunn's Fig. 6 with the horizontal axis shifted 6 dB; that is, what he calls 9 dB of clipping is plotted here as 3 dB below the epl.

In this study, the calculation of power loss is very close to Wathen-Dunn's and is somewhat a repeat of his work, except that the present study is a direct empirical calculation of the loss, rather than a theoretical calculation based on an empirical probability function.

IX. DIFFERENCES BETWEEN MALE AND FEMALE CONTINUOUS SPEECH

Some of the results of this study have been reported separately for male and female speech because there appears to be a real difference in the speech activity factor in the two sets of speech samples.

The average epls for men and women were adjusted to be roughly equal. The final average epls were measured as −5.86 dBm for women and −7.18 dBm for men. The women's levels therefore averaged 1.3 dB higher than the men's. However, measurements of average power (true rms) show that women averaged −16.46 dBm, and men averaged

−16.02 dBm. While the women's speech samples had higher levels, they had a lower average power than the men's.

This can be explained if during the editing process to produce continuous speech, the author allowed more short gaps in the women's speech than the men's, resulting in a lower activity factor for women. This is borne out by measurements of percent of A-to-D samples crossing a fixed −25 dBm threshold, which show an average of 35.9 percent for women, and 48.3 percent for men. This difference would cause the higher level women's speech to have less average power.

Once again, we are back to the problem of determining "when speech is present." Even though all editing was done in a two month period by the author, and even though the same telephone circuit, recording equipment, and editing equipment was used for both sets of data, consistent differences exist between the data for men and women. This finding should be a further warning that the results found here are intended only as guidelines and not as rigid specifications, since they can be influenced by the arbitrary design of the speech detection process.

X. RELATION OF PEAK CLIPPING TO VU

In an earlier unpublished study, the author suggested that VUs might be obtained from epls by subtracting 11 dB from the epl. This was based on extensive data from one highly trained observer, plus sample data from two other observers.

In the present study, the epl + 6 dB is suggested as a practical upper bound to the speech waveform. Thus, using the 11 dB epl to VU conversion, the expected VU would be about 17 dB below the highest peaks. This result is remarkably close to Noll's estimate of 17.2 dB as the peak-VU factor.[4]

It would appear that the data presented here might be applied to VU data by using the 11 dB conversion for epl to VU. A large body of VU data exists from field trials and the present results might be applied to these data. The author has also shown that VU data taken from many observers can be extremely erratic and at times can seem almost random.[3] Because of the variability in VU measurements, care should be taken in combining the results of the present study with existing VU data.

XI. SUMMARY

This study sought to examine speech signal statistics relevant to peak clipping and threshold crossings. All measurements were made on continuous speech. The following results were obtained:

(i) Instantaneous voltages rarely occur at levels higher than the epl + 6 dB.
(ii) The epl is exceeded by the instantaneous waveform about 2.5 percent of the time.
(iii) An additional 9 percent of time above threshold is gained for every 5 dB drop in threshold below the epl.
(iv) The average power in continuous speech is about 10 dB below the epl.
(v) The power in a "phantom noise signal", representing speech-correlated clipping noise, can be approximated by the curve in Fig. 4.
(vi) The power loss in continuous speech resulting from clipping can be approximated by the curve in Fig. 6.
(vii) VUs can be roughly approximated by subtracting 11 dB from epls. The other results might then be appropriately modified for VU data.

REFERENCES

1. Brady, P. T., "Equivalent Peak Level: A Threshold-Independent Speech-Level Measure," J. Acoust. Soc. Amer., *44*, No. 3 (September 1968), pp. 695–699.
2. Shearme, J. N., and Richards, D. L., "The Measurement of Speech Level," Post Office Elec. Eng. J., *47*, 1954, pp. 159–161.
3. Brady, P. T., "Need for Standardization in the Measurement of Speech Level," J. Acoust. Soc. Amer., *50*, No. 2 (August 1971), pp. 712–714.
4. Noll, A. M., unpublished work.
5. Brady, P. T., "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44*, No. 1 (January 1965), pp. 1–22.
6. Brady, P. T., "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., *47*, No. 1 (January 1968), pp. 73–91.
7. Brady, P. T., "A Statistical Basis for Objective Measurement of Speech Levels," B.S.T.J., *44*, No. 7 (September 1965), pp. 1453–1486.
8. Wathen-Dunn, W., and Lipke, D. W., "On the Power Gained by Clipping Speech in the Audio Band," J. Acous. Soc. Amer., *30*, No. 1 (January 1958), pp. 36–40.
9. Davenport, W. B., "An Experimental Study of Speech-Wave Probability Distributions," J. Acoust. Soc. Amer., *24*, No. 4 (July 1952), pp. 390–399.

Contributors to This Issue

JAMES L. BLUE, A.B., 1961, Occidental College; Ph.D., 1966, California Institute of Technology; Bell Laboratories, 1966—. Mr. Blue first worked on noise and avalanche diode oscillators. He has also done work in the field of numerical mathematics. Currently he is involved in the development of computer aids for integrated circuit design and testing. Member, American Physical Society, American Association for the Advancement of Science, IEEE, Phi Beta Kappa.

BARRY S. BOSIK, B.E.E., 1968, City College of New York; M.S.E.E., 1969, Princeton University; Bell Laboratories, 1968—. Mr. Bosik has worked on the transmission of analog *Picturephone*® signals over loop facilities. He is presently engaged in Ph.D. research at Columbia University concerning the spectral representation of digital coded signals. Member, Eta Kappa Nu, Tau Beta Pi.

PAUL T. BRADY, B.E.E., 1958, Rensselaer Polytechnic Institute; M.S.E.E., 1960, Massachusetts Institute of Technology; Ph.D., 1966, New York University; Bell Laboratories, 1961—. Mr. Brady has worked in modeling on-off speech patterns and speech-level distributions, especially as they occur in two-way conversation over circuits containing voice-operated devices and transmission delay. Member, Acoustical Society of America, Sigma Xi.

ALLEN GERSHO, B.S., 1960, Massachusetts Institute of Technology; M.S., 1961, and Ph.D., 1963, Cornell University; Bell Laboratories, 1963—. During the 1966–67 academic year, Mr. Gersho was Assistant Professor of Electrical Engineering at the City University of New York. He has performed research in time varying and nonlinear signal processing, synchronization, adaptive filtering, and the statistical approach to digital filter design.

JEAN-MARIE GOETHALS, M.S.E.E., 1961, and Ph.D., 1969, Louvain Catholic University, Belgium; MBLE Research Laboratory, Brussels, Belgium, 1963—. Mr. Goethals has been working on algebraic coding theory and applied combinatorial mathematics. He spent the Spring semester (1970) at the University of North Carolina, Chapel Hill, N. C., as a visiting lecturer. He is presently part-time lecturer at the Louvain

Catholic University, where he delivers courses on information theory and coding, and discrete mathematics. Member, A.M.S., IEEE, Société Mathématique de Belgique.

Nancy Y. Graham, B.A. (Mathematics), 1959, and M.A., 1962, University of California at Berkeley; Bell Laboratories, 1970—. Mrs. Graham has worked on problems in speech synthesis, plotting of perspectives, and linear programming. Member, American Mathematical Society, Sigma Xi.

Jessie MacWilliams (Mrs. F. J.), B.A., 1939, M.A., 1941, Cambridge University (England); Ph.D., 1962, Harvard University; Bell Laboratories, 1956—. Mrs. MacWilliams has worked in transmission networks development and data communications engineering, and is now in the Mathematics and Statistics Research Center. Member, Mathematical Association of America, American Mathematical Society.

R. F. W. Pease, B.A., 1960, M.A. and Ph.D., 1964, University of Cambridge; Bell Laboratories, 1967—. Mr. Pease held a faculty appointment at the University of California at Berkeley prior to joining Bell Laboratories and worked on electron microscopy. At Bell Laboratories he has worked on the digital encoding of television signals. Presently, he is engaged in using electron beams to make integrated circuits.

George C. Reis, B.S.E.E., 1959, Milwaukee School of Engineering; M.S.E.E., 1962, Drexel University; Ph.D., 1967, University of Pennsylvania; RCA, 1959–1962; Drexel University, 1962–1967; Bellcomm, 1967–1971; Bell Laboratories, 1971—. Mr. Reis is presently involved in error evaluation of digital data systems. Member, Eta Kappa Nu, Tau Beta Pi, IEEE.

Neil J. A. Sloane, B.E.E., 1959, and B.A. (Hons.), 1960, University of Melbourne, Australia; Postmaster General's Department, Commonwealth of Australia, 1956–1961; M.S., 1964, and Ph.D., 1967, Cornell University; assistant professor of electrical engineering, Cornell University, 1967–1969; Bell Laboratories, 1969—. Mr. Sloane is engaged in research in coding theory, communication theory, and combinatorial mathematics. Member, IEEE, American Mathematical Society, Mathematical Association of America.

ALAN N. WILLSON, JR., B.E.E., 1961, Georgia Institute of Technology; M.S.E.E., 1965, and Ph.D., 1967, Syracuse University; International Business Machines Corporation, 1961–1964; Bell Laboratories, 1967—. Mr. Willson has been engaged in research concerning the stability of distributed circuits, properties of nonlinear networks, and digital filters. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi, Society for Industrial and Applied Mathematics.

B. S. T. J. BRIEF

Proof of a Convexity Property of the Erlang B Formula

By E. J. MESSERLI

(Manuscript received January 5, 1972)

I. INTRODUCTION

Consider Poisson traffic offered to a group of n trunks. Blocked calls are cleared, and call holding times are independent and identically distributed. In equilibrium, the call congestion is given by the well-known Erlang B formula[1]

$$B(n, a) = \frac{\frac{a^n}{n!}}{\sum_{k=0}^{n} \frac{a^k}{k!}},$$

where a is the offered load in erlangs. We prove here that

$$B(n, a) - B(n + 1, a) < B(n - 1, a) - B(n, a) \qquad n = 1, 2, \cdots, \qquad (1)$$

i.e., strict convexity with respect to the number of trunks.† For a trunk group with sequential assignment of offered calls, these inequalities simply state that the load carried on the last trunk is monotonically decreasing with the number of trunks–a commonly accepted fact which is basic to economic alternate routing in network engineering.[2] Nevertheless, analytical verification of (1) has apparently not been published. The proof given here offers one approach to verifying convexity properties for other loss systems as well.

II. DEVELOPMENT OF MAIN RESULT

For an m-trunk group with Poisson offered traffic, consider a class of operating policies of the form:

† Syski[1] mentions some results for the analytic continuation of $B(\cdot, a)$, but convexity on the integers does not seem to follow.

If a call arrives when there are i calls in progress, $i < m$, accept the call with probability δ_i, where δ_i is fixed.

A policy may be represented by a vector $\delta(m) = (\delta_0, \delta_1, \cdots, \delta_{m-1})$. For any policy $\delta(m)$, identifying states with the number of calls in progress, a stationary solution to the birth-and-death equations exists[1] with carried load given by:

$$c(\delta(m)) = \sum_{k=1}^{m} k \frac{a^k}{k!} \delta_0 \cdots \delta_{k-1} \left(1 + \sum_{k=1}^{m} \frac{a^k}{k!} \delta_0 \cdots \delta_{k-1}\right)^{-1}.$$

Lemma 1: The carried load $c(\delta(m))$ is maximized by the unique optimal policy $1(m)$ defined by $\delta_i = 1$, $i = 0, 1, \cdots, m - 1$, i.e., if there is an empty trunk, accept the call.

This result is obvious and a proof is straightforward: if i is the first index such that $\delta_i < 1$ for $\delta(m)$, it is easy to show that

$$\frac{\partial c}{\partial \delta_i}(\delta(m)) > 0.$$

The result is also implicit in the fundamental inequalities developed by Beneš.[3] Several proofs for the lemma were later given in Ref. 4.

For a group of n trunks, with sequential assignment, let the carried load on the ith trunk be given by $a_i = a_i(a)$, $a > 0$. Since $a_1 + a_2 + \cdots + a_n = a(1 - B(n, a))$, the inequalities (1) hold if and only if

$$a_1 > a_2 > a_3 \cdots .$$

The main result is:

Theorem 1: The sequence $a_1, a_2, \cdots$, where $a_i = a(B(i - 1, a) - B(i, a))$, satisfies $a_1 > a_2 > \cdots$ for any positive offered load a.

Proof: Suppose not, i.e., that for some n and $a > 0$, $a_{n-1} \leqq a_n$. We shall show that this leads to a contradiction of Lemma 1. Thus, consider a group of $n - 1$ trunks with calls placed according to:

(*i*) Sequential assignment for the first $n - 2$ trunks.

(*ii*) Overflow calls from the first $n - 2$ trunks are offered to the last trunk according to the status of a dummy trunk. If the dummy trunk is free, the call is rejected, and a dummy call with the same holding time distribution is placed on the dummy trunk. If the dummy trunk is busy, the call is offered to the last trunk.

For this system, define

$$\bar{P}_i = \text{Prob (call put up} \mid i \text{ real calls in progress}).$$

It is easy to see that $\bar{P}_i = 1$, $i = 0, 1, \cdots, n-3$, and that $\bar{P}_{n-2} < 1$. Moreover, the carried load is equal to that for an $n-1$ trunk group system corresponding to the policy $\delta^*(n-1) = (\bar{P}_0, \cdots, \bar{P}_{n-2})$. But the carried load $c(\delta^*(n-1))$ is given by $a_1 + a_2 + \cdots + a_{n-2} + a_n$. Since $a_{n-1} \leqq a_n$,

$$c(\delta^*(n-1)) \geqq a_1 + a_2 + \cdots + a_{n-1} = c(1(n-1))$$

which contradicts Lemma 1. This completes the proof.

Remark: Subsequent to the appearance of this development in unpublished form, other proofs have been given. Krupp[5] gives an algebraic proof. Descloux[6] gives both an algebraic proof, and a proof of the equivalent result to (1) for renewal input. Buchner and Neal[7] also give a proof for the generalization to renewal input. A reviewer pointed out that the result can be proved by comparing occupancy probabilities on the nth trunk for sequential assignment, and for sequential assignment modified so that an overflow call from the first $n-2$ trunks chooses one of the last two trunks equally likely if both are free.

REFERENCES

1. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Edinburgh and London: Oliver and Boyd, 1960.
2. Truitt, C. J., "Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks," B.S.T.J., *33*, No. 2 (March 1954), pp. 277–302.
3. Beneš, V. E., "Some Inequalities in the Theory of Telephone Traffic," B.S.T.J., *44*, No. 9 (November 1965), pp. 1941–1975.
4. Crawford, J. W., "General Solution of a Class of Stochastic Single Commodity Network Flows," unpublished work, 1969.
5. Krupp, R. S., "Convexity of Erlang B and C," unpublished work, February 1972.
6. Descloux, A., "On the Convexity of the Erlang Loss-Function," unpublished work, February 1972.
7. Buchner, M. M., Jr., and Neal, S. R., "Monotonicity of the Load on the Last Server," unpublished work, February 1972.