1971

Peak harmonic distortion due to quantization

On-Ching Yue

1971

Peak harmonic distortion due to quantization

On-Ching Yue

PEAK HARMONIC DISTORTION DUE TO QUANTIZATION

by

On-Ching Yue

A Thesis Submitted

in

Partial Fulfillment

of the

Requirements for the Degree of

MASTER OF SCIENCE

in

Electrical Engineering

Approved by:

Prof. Edward R. Salem
(Thesis Advisor)

Prof. Harvey E. Rhody

Prof. W. F. Walker

Prof. W. F. Walker
(Department Head)

DEPARTMENT OF ELECTRICAL ENGINEERING

COLLEGE OF APPLIED SCIENCE

ROCHESTER INSTITUTE OF TECHNOLOGY

ROCHESTER, NEW YORK

December, 1971

ACKNOWLEDGEMENT

The author wishes to thank Professor E. Salem for his guidance in this thesis, and Dr. G. J. Sehn for many stimulating discussions on the subject matter.

ABSTRACT

In this thesis, the problem of quantization noise is presented, and recent efforts in this area are reviewed. With the motivation for further investigation into the problem explained, the purpose of this thesis is stated to be the determination of peak harmonic distortion due to quantization for predominantly single frequency inputs. Two cases were examined with pure sinusoid and sine wave plus Gaussian bandlimited white noise. The method used was to simulate the quantization process on the computer, and to use a Fast Fourier Transform algorithm to analyze the spectra of the quantized signals.

For pure sinusoidal inputs, the location of the peak harmonic distortion in the quantization noise spectrum was found to be very sensitive to the degree of loading of the quantizer. However the magnitude of the peak distortion when plotted as a function of the number of bits totally used by the input was fitted very well with a straight line of slope -6dB/bit. Moreover the largest component in the quantization noise spectrum was observed to be about 4dB above the average noise spectral density across the entire frequency band of observation. The addition of noise to the sine wave was anticipated to have a smoothing effect on the quantization noise spectrum. This phenomenon was observed for a specific set of input noise samples but the results are not conclusive, becuase after further investigation of the noise generation mechanism, the statistical properties of the synthesized noise signal were found to be unsuitable for analysis of power spectra.

However since the pure sinusoidal inputs represent the worst case condition for harmonic distortion due to quantization, the measured peaks

will provide the upper bounds necessary for specifications in engineering system designs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

In recent years it has become increasingly popular to use digital techniques to process analog signals. In the field of communication, digital modulation schemes, for example pulse code modulation (PCM), are preferable over analog methods, for example amplitude modulation (AM), where the performance index is measured in terms of error rate and the efficient use of the available channel capacity [1]. With the advent of high speed and low priced computers, more and more processes, in chemical plants for example, employ direct digital control (DDC), for better error performance in the systems [2]. Moreover in the field of filtering, which is applicable to both communication and automatic control, digital filters are becoming popular because of the accurate control over the transfer functions and the attainability of sharp skirt characteristics which are beyond the reach of analog filters.

However between the domain of logic circuits and the real world of continuous signals which may be voice, or sonar echoes, or system parameters being monitored by sensors, there is a vital link which consists of a sampler and a quantizer. The effect of sampling is well known via Shannon's sampling theorem [3], so this thesis is primarily concerned with the "damage" done to the original signal by the quantizer.

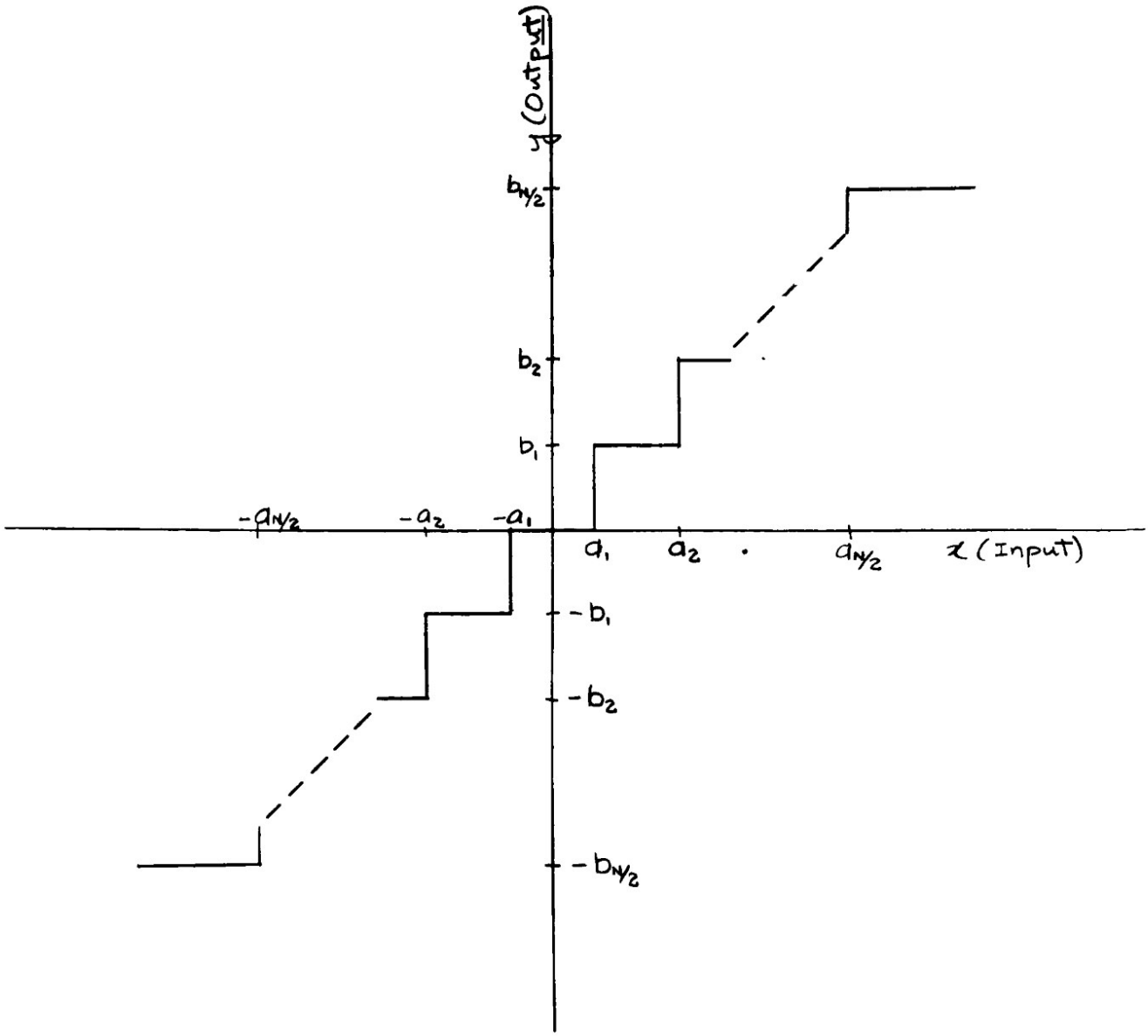
A quantizer is a nonlinear device which produces the same output values for all input amplitudes which lie in each of a finite number of amplitude ranges. This kind of a "stair-case" transfer function is illustrated in figure 1. The error signal which is given by the difference between the input and output signals of the quantizer is commonly referred to as quantization noise. As we shall see in the next section, the subject of

quantization noise or more specifically, the noise spectrum, has been of considerable interest to engineers both in the communication and control fields, and consequently a fair amount of study has already been done in this area. Therefore the question arises: why should there be another investigation of the spectrum of quantized signals.

The motivation of this thesis was derived from an engineering problem in sonar applications. The problem deals with the measurement of harmonic distortion due to sonar transducers. Although transducers are often used as linear devices, their linear transfer characteristics are really only first order approximations to nonlinear functions, sufficiently accurate for small signals. Consequently when the transducer is used as a projector, operating at high power, the acoustic signal transmitted into the medium will contain many harmonics due to the nonlinearity of the transfer function, even though the input signal may be a pure sinusoidal. In order to measure the harmonic distortion introduced by the projector, the acoustic signal is received by another hydrophone and harmonic analysis is performed. If the signal processing and data transmission are to be done digitally, then the signal received at the output of the hydrophone will have to be sampled and quantized. The question arises: how many bits should be used in the quantizer, such that the quantization noise will not contribute any significant inaccuracy to the determination of harmonic distortion which would be performed after additional signal processing. This information is important because it affects the design of the special-purpose hardware for signal processing and the bit rate for data transmission.

Since the quantization noise spectra for predominantly single frequency inputs are very irregular and each harmonic amplitude varies widely depending on the loading of the quantizer, it is more important for our purposes to know the upper bound of the noise spectrum, regardless of where the maximum spectral density may be located. Therefore the purpose of this thesis is to determine the peak harmonic distortion due quantization as a function of the number of bits in the quantizer. This relationship will enable an engineer to determine the optimal size of the quantizer used in the system to meet a set of accuracy specifications subject to such contraints as cost and bit rate.

After a survey of the literature on the problem of quantization it was decided that the best approach would be experimental - using computer simulation, and a Fast Fourier Transform (FFT) algorithm to perform the spectral analysis. Before the experimental procedures and results are described, a review of the recent efforts in the area of quantization is presented in the next section, to justify the choice of approach in this thesis.



$$y = \begin{cases} 0 & \text{for } |x| < a_1 \\ \pm b_k & \text{for } a_k \leq \pm x \leq a_{k+1}, \quad 1 \leq k < N/2 \\ \pm b_{N/2} & \text{for } a_{N/2} \leq \pm x \end{cases}$$

Transfer Function of Quantizer (N-level).

Figure 1

II. HISTORICAL BACKGROUND

As soon as one begins a literature survey on the subject of quantization noise, one discovers that the work of Bennett[4] is the reference quoted by almost all the authors. In 1948 Bennett published his investigation of the spectra of quantized signals for uniformly spaced quantizers, which means, in terms of figure 1, $a_{k+1}-a_k = 2a_1$ and $b_{k+1}-b_k = b_1$ for $1 \leq k < N/2$.

Bennett derived a simple but extremely useful expression for the mean square error of quantization or the noise power. He observed that for input signals which have amplitude variations much larger than the quantum step, $b_1 = 2a_1$, the error signal resembles a saw-tooth waveform, going from $-a_1$ to $+a_1$ with arbitrary slope. Exceptions occur when the slope of the input signal changes sign within a quantum step resulting in large deviation from a saw-tooth, but with the assumption of large rms value for the input signal as compared to the quantum step, these exceptions are rare. The mean square value of a saw-tooth function with arbitrary slope and bounded by $\pm a_1$ is easily calculated. The equation of a typical line segment of the function is :

$$e_s(t) = st, \qquad \text{for } -a_1/s < t < +a_1/s$$

where s is the slope and t is arbitrarily referenced to the midpoint of the segment. Then the mean square error for this segment is

$$\begin{aligned}\overline{e_s^2(t)} &= (s/2a_1)\int_{-a_1/s}^{a_1/s} e_s^2(t)\,dt \\ &= (s/2a_1)(t^3/3)\Big|_{-a_1/s}^{a_1/s} \\ &= (2a_1)^2/12 \end{aligned} \tag{2.1}$$

Since $\overline{e_s^2(t)}$ is independent of the slope s, then the calculated mean square error of $(2a_1)^2/12$ is true for the entire error spectrum.

Panter[5] derived a more general expression for the mean square error (which reduces to equation(2.1)), from a statistical point of view. The assumption of uniform step size is removed but unity slope is still maintained in the transfer function, i.e.

$$b_k = (a_{k+1}+a_k)/2 \qquad \text{for } 1 \leq k < N/2.$$

Let p(x) be the probability density function of the input signal. Then the mean square error voltage $\langle (x-b_k)^2 \rangle$ associated with the quantization of the input signal assigned to a_k is given by

$$\langle (x-b_k)^2 \rangle = \int_{a_k}^{a_{k+1}} (x-b_k)^2 p(x)dx \qquad (2.2)$$

Panter made an assumption that the probability density function is effectively constant within each step, but allowed it to vary from step to step. This assumption is basically the same as the one made by Bennett with regard to sufficiently small quantum step relative to signal variations. Then it follows that equation (2.2) can be evaluated as:

$$\langle (x-b_k)^2 \rangle = p_k \int_{a_k}^{a_{k+1}} (x-b_k)^2 dx$$

$$= (d_k)^3 p_k/12$$

where $d_k = a_{k+1} - a_k$

and $p_k = \frac{1}{d_k} \int_{a_k}^{a_{k+1}} p(x)dx.$

Inother words p_k is the average probability density for $a_k \leq x \leq a_{k+1}$ and $(p_k d_k)$ is the probability that the input signal lies in the same range.

The overall mean square error $\langle e^2 \rangle$, which is the total noise power due to quantization, is the sum of the mean square errors introduced at each quantized levels. Therefore

$$\langle e^2 \rangle = \sum_k \langle (x-b_k)^2 \rangle = (1/12) \sum_k d_k^3 p_k$$

$$= (1/12) \sum_k d_k^2 (p_k d_k)$$

$$= \langle d_k^2 \rangle /12 \qquad (2.3)$$

where $\langle d_k^2 \rangle$ is the weighted average of the square of quantum step sizes. Note that for uniform quantizers, $d_k = d$ is a constant yielding

$$\langle e^2 \rangle = d^2/12 \qquad (2.4)$$

which is the same as equation(2.1), except that the latter is a time average while equation(2.4) is an ensemble average.

Equation(2.1) is quite accurate as Bennett had observed experimentally and generally provides sufficient information to determine the degradation of system performance due to quantization in terms of signal-to-noise ratios. However if one is interested in more details about the spectrum of the quantization noise , then further analysis is necessary.

Analyses of nonlinear problems are usually mathematically tedious and yield theoretical results which, in general, are not easily interpreted physically. Therefore in order to get meaningful results, the analyses have to be restricted in their scope, so that approximations can be applied. In the case of the quantizer, which has a nonlinear transfer function, the output spectra of only a limited class of input

signals are analyzed.

Bennett observed that with single or double frequency signals, the noise spectra are "ragged", and the amplitudes of the harmonics generated oscillate violently with the input magnitudes. Consequently he chose to use narrowband Gaussian white noise as the input. Since both the autocorrelation function and the probability density function of the input signal are well defined, the autocorrelation function, and therefore the power spectrum, of the quantized error at the output can be calculated in a straightforward manner. However we shall only outline the approach, without the detailed mathematical manipulations.

Assuming the quantization noise to be stationary and ergodic, the autocorrelation function of the output error is, by definition, given by:

$$\begin{aligned} R_e(v) &= \overline{e(t)e(t+v)} \\ &= \langle e(t)e(t+v)\rangle \\ &= \langle (x_1-md)(x_2-nd)\rangle \qquad \text{for } m,n = 0,\pm1,\pm2,\ldots \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-md)(x_2-nd)p(x_1,x_2)dx_1dx_2 \end{aligned} \tag{2.5}$$

where x_1 and x_2 are the amplitudes of the input signals at time t and $t+v$; md and nd are the output levels assigned by the quantizer to x_1 and x_2 respectively; d is the uniform quantium step; and $p(x_1,x_2)$ is the two-dimensional probability density function which is Gaussian. The correlation coefficient, φ, is given by

$$\varphi(v) = R_i(v)/R_i(0) \tag{2.6}$$

where $R_i(t)$ is the autocorrelation function of the input signal, and since the input spectrum is assumed to be narrowband and flat, $R_i(t)$ is a "sinc" function. Again under the assumption of small quantum

step size as compared to input rms, i.e. $d^2 \ll R_i(0)$, equation(2.5) can be solved and approximated as

$$R_e(v) \cong (d^2/2\pi^2)\sum_{n=1}^{\infty}(1/n^2)\exp\,(-4n^2\pi^2[1-\varphi(v)]/k) \qquad (2.7)$$

where k is $d^2/R_i(0)$. The power spectrum of the quantization noise $S_e(f)$ is given by taking the Fourier Transform of $R_e(v)$.

$$S_e(f) = (k/2\pi^2)(3k/2\pi)^{1/2}B(3kf^2/2) \qquad (2.8)$$

where $B(z) = \sum_{n=1}^{\infty}(1/n^3)\exp\,(-z/n^2)$.

Bennett measured the power spectrum of quantization noise for 5, 6 and 7-bit quantizers, with inputs of thermal noise and a 16-tone signal, and found good agreement with the spectrum predicted by equation(2.8).

Nonlinear problems are quite common in science and engineering, and one of the standard techniques to solve these problems is the "transform method"[6], where the nonlinear characteristics are Fourier transformed to facilitate analyses. Banta[7] and Rowe[8] applied this method to uniform quantizer where the error characteristics, as a function of the input amplitude, is periodic. Actually since any real quantizer only has a finite number of bits, the error function is of finite duration and therefore not periodic. However if we assume that the input signal would never overload or saturate the quantizer, then equivalently there are infinite number of steps in the quantizer, and the error characteristic is periodic. In Bennett's analysis, this

assumption of an infinite size was also used and in his experiments, he observed that the peak voltage of thermal noise never exceeded appreciably four times its rms voltage. Since a Gaussian probability function is assumed for the amplitude of the noise, the probability of the instantaneous voltage exceeding four times the standard deviation is less than 10^{-4}. Therefore Bennett was able to avoid saturating the quantizer.

The error characteristic of an infinite, uniform quantizer, as given by the difference between the output and the input amplitudes, can be expanded as a Fourier series in x, the input.

$$e(x) = (1/d)\sum_{n=-\infty}^{\infty} c_n \exp(j2\pi n f_0 x) \qquad (2.9)$$

where $f_0 = 1/d$,

and $$c_n = \int_{-d/2}^{d/2} e(x) \exp(-j2\pi n f_0 x)dx$$

integrating by parts, $u = e(x) = x$, and $dv = \exp(-j2\pi n f_0 x)dx$,

$$= [x \exp(-j2\pi n f_0 x)\Big|_{-d/2}^{d/2} - \int_{-d/2}^{d/2} \exp(-j2\pi n f_0 x)dx]/(-j2\pi n f_0)$$

$$= (d/2)[\exp(-j2\pi n f_0 d/2) + \exp(j2\pi n f_0 d/2) - 0]/(-j2\pi n f_0)$$

$$= (jd^2/4\pi n)[\exp(-jn\pi) + \exp(jn\pi)]$$

$$= (jd^2/2\pi)(-1)^n/n$$

$$c_0 = 0$$

Then for any input signal $x(t)$, the error signal is given by

$$e(t) = (jd/2\pi) \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} ((-1)^n/n) \exp(j2\pi n x(t)/d) \qquad (2.10)$$

Rowe solved equation(2.10) for a pure sinusoidal input,

$$e(t) = (jd/2\pi) \sum_{n=-\infty}^{\infty} ((-1)^n/n) \exp(j2\pi n A \sin 2\pi f t)/d$$

$$= (-2d/\pi) \sum_{\substack{k=1 \\ k\text{-odd}}}^{\infty} \sin 2\pi k f t \left[\sum_{n=1}^{\infty} ((-1)^n/n) J_k(2\pi n A/d) \right] \qquad (2.11)$$

where $\{J_k(z)\}$ are Bessel functions of the first kind.
For bandlimited Gaussian noise input, Rowe obtained the same result as Bennett for the autocorrelation function of the quantization noise, which reduces to equation(2.7) under the same conditions stated before.

Banta went one step further and used a combination of a deterministic signal and Gaussian noise as input. The result obtained for the autocorrelation function of the output noise due to quantization is exact and general, except for the assumption of statistical independence between the signal and the noise. However the expression is extremely difficult to use and one special case was treated, where there was no signal component and the input noise autocorrelation function is triangular. After finding the quantization noise spectrum to be bounded by a certain value, Banta concluded that a low-pass filter with bandwidth comparable to the main lobe of the input noise spectrum will effectively suppress the quantization noise relative to the input. This same conclusion could probably be arrived at by

observing that the quantization noise is broadband as compared to the input spectrum, because of the cross-modulation of the input frequency components due to the nonlinearity. Consequently the best way to recover the input signal is by low pass filtering.

The analyses of quantization noise discussed so far are limited to uniform quantum steps only, but there are some practical situations where nonuniform quantizers are more desirable. Bennett did some experiments with tapered quantizers, because in voice transmission it is advantageous to emphasis the weak signal components. Max[9] analyzed a more interesting problem. He defined the distortion of quantization to be the expected value of some function of the error between the input and the output. Then for a fixed number of quantum steps and a given input probability density function, the step sizes are chosen to minimize the distortion, allowing the quantizer to be saturated. The equations cannot be solved analytically, so Max tabulated some numerical results for a Gaussian noise input. Algazi[10] used an even error weighting function (mean square, for example) for distortion measure and provided some useful approximations to the equations given by Max. Hurd[11] also used the transform method to derive the output autocorrelation of a nonuniform quantizer, when the input is a sine wave plus stationary Gaussian noise.

Returning to the problem stated in the previous section, where we are interested in uniform quantizer, we can see that one approach to solve the problem is to use equation(2.10) and for different kinds of input signals, the autocorrelation and the power spectrum of quantization noise can be calculated. However in view of the complexity

of equations(2.10) and (2.11), it is anticipated that numerical methods have to be used, in which case, we feel that an experimental approach would be much simpler. Since spectral analysis can be performed quickly and economically by the computer, using the Fast Fourier transform (FFT) algorithm, the whole system was simulated on the computer.

III. DESCRIPTION OF COMPUTER SIMULATION

In this section we shall describe the various aspects of the computer simulation as shown in figure 2: the generation of input signals, the quantizer and the Fourier transformer.

As it had been mentioned in Section I, the motivation for this investigation is the measurement of harmonic contents in the output of an acoustic transducer for purely sinusoidal inputs. As engineers are constantly making the transfer function of the transducer more and more linear which means less harmonic distortion, rather than introducing an arbitrary amount of harmonic contents in the simulated acoustic signals, we have restricted our study to quantizer input signals which have only a single frequency component. However to simulate the actual signals received by the measuring hydrophone , which will be be contaminated with ambient noise, bandlimited white Gaussian noise is also added to the pure sinusoid. Therefore two kinds of signals were used as inputs to the quantizer:pure sine wave, and fundamental plus noise. The generation of a pure sinusoid can easily be done by the computer but the generation of the noise is more complicated. One approach is to use an IBM scientific subroutine which generates random numbers with an approximately Gaussian distribution. However since we prefer better control over the noise spectrum, the noise signal is first synthesized in the frequency domain and then Fourier transformed into time. The noise power spectrum is chosen to be flat and have bandwidth of about half the sampling frequency. The randomness is introduced when the power spectrum is converted into a frequency spectrum by choosing the phase angles of the different frequency components from a random number generating subroutine with uniform

probability density function from zero to 2π. This synthesized noise signal has 63 frequency components and is shown in Appendix A to approximate Gaussian noise very well.

The input noise power of the fundamental plus noise signal is chosen to be comparable to the noise power due to quantization, because if the input noise level is much higher than that of the quantizing noise , then the input rather than the quantizing noise would place a limitation on the accuracy of the harmonic measurements.

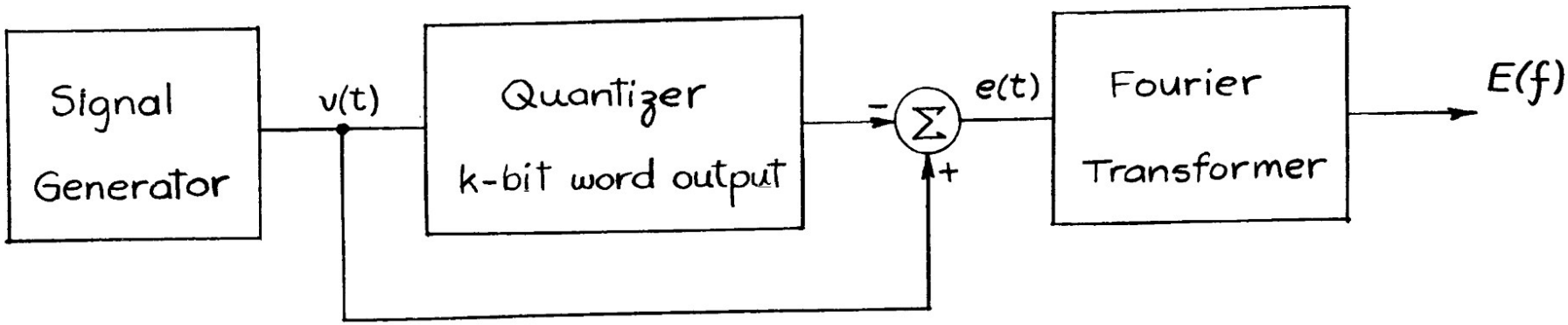
The quantization process is quite similar to the truncation of floating point to fixed point numbers in the computer. However since the computer truncates the absolute value of a number and then puts back the sign, for numbers which range from negative to positive, the transfer characteristic of this truncation process has a step size at zero output twice as large as at any other level . To avoid this nonuniform quantizing region, the subroutine which simulates a quantizer first linearly translates all input values to the positive domain, performs the truncation and then centers all values back around zero.

The number of bits in a quantizer can be very large, but with the contraints of signal- to-noise ratioes in the real world, a 12-bit quantizer is quite adequate. Moreover the entire dynamic range of the quantizer is sometimes not utilized, so that the quantizer has effectively an even smaller number of bits. For our purposes, we will use quantizer sizes ranging from 3- to 12-bit, where the sign is also included. In practice to avoid saturation of the quantizer the input amplitude is estimated a priori and the quantizer is only loaded half way by the predicted value. Similarly in the simulation the quantizer is

loaded half way by the amplitude of the fundamental component, and the effect of small changes in loading (within one quantum step) is also investigated.

The numerical counterpart of the continuous Fourier transform is the discrete Fourier transform (DFT) operating on denumerably infinite sets which can be interpreted as time and frequency samples. However since the computer can only work with finite data blocks, we must assume that the continuous signal is periodic and bandlimited so that only a finite number of samples within a period is sufficient to describe the signal completely. The use of the DFT to analyze signals not satisfying the above assumption would result in truncation and aliasing errors[12]. However since the signals used in the simulation are both bandlimited and periodic (including the noise), the noise spectra generated are exact. The subroutine used to perform the Fourier transform is a radix-2 fast Fourier transform(FFT) algorithm.

The input signal is put through the quantizer and the error, which is the difference between input and output, is Fourier transformed. Another subroutine converts the Fourier coefficients (complex) into magnitude in dB referenced to the amplitude of the fundamental component and the peak harmonic error is searched. One final comment on the simulation is that word size of the computer used in this experiment was 32 bits, so that the roundoff (quantization) errors introduced by the machine in computation were negligible as compared to the measured quantization noise for word sizes of 12 bits and less.



Block Diagram of Computer Simulation

Figure 2

IV. DISCUSSION OF TEST RESULTS

Throughout the simulation, a data block size of 128 points was used and the dynamic range of the spectrum analyzer (FFT subroutine) was more than 110dB, which also included the effect of the impurities of the sine wave generation subroutine in the computer.

Single frequency signals of f_o, $3f_o$, $9f_o$, $15f_o$, $21f_o$, $33f_o$ and $63f_o$ were used as inputs, where f_o is equal to 1/128 of the sampling frequency. The number of bits in the quantizer occupied by the input signal ranged from 3 to 12, sign included, and the peak input amplitudes were scaled to equal the maximum quantizer output for the chosen number of bits (full loading). To demonstrate the effect of slight changes in loading, the quantizer was also down loaded to 20% of a quantum step from the maximum.

Figures 3 and 4 show the quantization noise spectra for input frequency of f_o and 3-, 6-, 9- and 12-bit quantizers. Since the inputs were pure sinusoids, the output noise spectra consisted of spectral lines which were odd harmonics of the input frequency (see equation (2.11)). However for the purpose of comparison between different spectra, in figures 3 and 4, straight lines were drawn between the spectral peaks, neglecting all the even harmonics. The spectral density was plotted on a dB scale referenced to the input amplitude.

From figures 3 and 4, we observe that the magnitudes of the spectral lines are very irregular, and only the 3-bit curve shows a decreasing trend for increasing frequency. For the other curves the spectral lines have local maxima which are sometimes separated by less than 1dB. On the same curves, the effect of down loading by 20% of a quantum step is also demonstrated. For a fixed number of bits,

we observe that a slight change of loading resulted in a completely different quantization noise spectrum. Of most significance to this thesis is the observation that the location of the highest spectral line (I within the bandwidth of half the sampling frequency) is very sensitive to the loading of the quantizer and no apparent functional relationship can be observed.

It should be pointed out that although the ratio of sampling to signal frequency was 128 in figures 3 and 4, the effect of aliasing, or the folding of the noise spectrum into a bandwidth of half the sampling frequency, is unavoidable. Therefore the curves shown are already results of interactions between different spectral lines, reinforcements or cancellations depending the relative phases. In practice the sampling frequency is probably not even rationally related to the input frequency, so that there would not be any interaction. However as Bennett had observed, the spectrum of the unsampled noise is very irregular and decays very slowly, so that the probability of having two strong spectral lines adding almost algebraically to yield a spectral line higher than the true peak without aliasing should be small. Consequently as long as we make the ratio of the sampling to signal frequency rational rather than integral, keeping the interaction of the spectral lines to a minimum, our estimate of peak harmonic distortion would be accurate. This hypothesis is observed to be true in the invariancy of the magnitude of the peak harmonic distortion to changes in the input frequency.

Table I shows both the location and the magnitude of the observed peak spectral lines in the quantization noise spectra for different combinations of input frequency and number of bits. First

we observe that the magnitude of the peak distortion is independent of the input frequency, for number of bits ranging from 3 to 8, and for larger number of bits, the separation is as much as 2.5dB for different input frequencies. Next we observe that although the location of the peak harmonic distortion is unpredictable as a function of the number of bits, the location for the different input frequencies can be calculated from the one at f_α by multiplying by the appropriate ratio and modulo 128 reduction. For example in the '6-bit' column of Table Ib, the location for M=3 can be obtained by $|(27\text{x}3) \bmod 128| = |81 \bmod 128| = 47$. Similarly for M=9, $|(47\text{x}3) \bmod 128| = |141 \bmod 128| = 13$, and M=15 $|(47\text{x}5) \bmod 128| = |235 \bmod 128| = 21$. The reason for taking the absolute value is that we are only observing in the bandwidth between 0 and 64f . The exception to the stated rule is in the '12-bit' column of Table Ib.

The reason for the two observed phenomena of regularity can easily be explained based on the assumption that the location and magnitude of the peak harmonic distortion are almost unaffected by sampling. We recall that the quantizer is a nonlinear device which means the output is dependent on the input amplitude, but it is also frequency independent. Therefore for the same degree of loading by different pure sinusoids, without sampling the quantization noise spectra are identical except for a scale factor in frequency, i.e. the magnitude and the harmonic number of the largest spectral line are identical. Assuming little interaction between spectral lines, the largest spectral line will remain to be the maximum, and the effect of sampling is only to bring the line into the basic bandwidth. As a result the location can easily be calculated. The exceptions for larger number of bits can be attributed

to the slower decay of the quantization noise spectrum, i.e. the noise becomes less correlated, resulting in a wider bandwidth.

The most significant result of this thesis is observed when the peak harmonic distortion is plotted as a function of number of bits in figure 5 for input frequency of $3f_o$. The experimental data can be fitted very well with a straight line of slope -6dB per bit. Table II tabulates the wideband quantization noise power and the results for 3f are plotted in figure 6 along with the theoretical noise power for white noise input ($d^2/12$). The fit of total power data by a -6dB/bit line can be expected by reasoning that by throwing away one bit, the available voltage aperture is halved and therefore the noise power will effectively increase by 6dB. However there is no apparent reason for the peak harmonic distortion to fall off approximately at -6dB/bit, especially considering the randomness of the location of the peak. This result is significant because it gives a simple relationship between the peak harmonic distortion and the average spectral density (a difference of about 4dB), which is independent of frequency and the size of the quantizer. It should be noted that in order to obtain the spectral density from figure 6, it is necessary to subtract the bandwidth.

For the inputs of pure sinusoid plus noise, the noise power had been chosen to be comparable to the quantization noise power due to the sinusoid alone. For a specific set of noise samples used in the computer simulation, a phenomenon as illustrated in figure 7 is observed. In other words the effect of introducing input noise is a reduction of the peak distortion with respect to the average quantization power, a smoothing effect. The crossover point is around $N_i/N_q=1$, where N_i and N_q are the

input and quantization noise powers repectively. This phenomenon can be expected because as the root-mean-square value of the input noise becomes comparable to quantum step size, the relationship between the quantization noise and the input sinusoid is masked and no particular harmonic can be expected to dominate the output noise spectra.

Since in any simulation of a probabilistic system, it is necessary to determine the variance of the measured parameters, power spectra in this case, which are only estimates of the true values, we ran the simulation several times with different sets of noise samples. The output noise spectra were quite different and the increase of block size to 512 did not result in any sign of convergence to one another. Consequently the noise generating mechanism was investigated in greater details and we discovered that the proposed synthesized noise signal is ergodic only up to the second moment, as shown in Appendix B. This means that the third and higher statistical characteristics of the time-samples generated are dependent on the specific set of phase angles chosen and do not agree with one another. Since the variance of the estimate of the autocorrelation function (power spectra) is dependent on the fourth moment of the noise, our observation about the quantization noise spectra for pure sinusoid plus bandlimited white Gaussian noise cannot be generalized.

Magnitude of Peak Harmonic Distortion (dB) / Input Frequency* \ No. of Bits	3	4	5	6	7	8	9	10	11	12
1	-26.07	-33.09	-38.37	-46.12	-52.51	-57.46	-66.05	-69.16	-75.34	-82.65
3	-26.07	-33.09	-38.37	-46.12	-52.51	-57.46	-66.04	-69.16	-75.33	-82.63
9	-26.07	-33.09	-38.37	-46.12	-52.51	-57.46	-66.07	-69.15	-75.35	-83.12
15	-26.07	-33.09	-38.37	-46.12	-52.50	-57.45	-66.75	-69.97	-76.11	-83.53
21	-26.07	-33.09	-38.37	-46.12	-52.51	-57.46	-66.69	-70.86	-76.64	-82.90
33	-26.07	-33.09	-38.37	-46.12	-52.51	-57.45	-66.74	-70.83	-76.36	-83.92
63	-26.07	-33.09	-38.37	-46.12	-52.49	-57.44	-66.45	-71.07	-78.67	-85.06

*Frequency normalization factor = sampling frequency/128

Table Ia. Magnitude of Peak Harmonic Distortion

Frequency of Peak Harmonic Distortion / Input Frequency* \ No. of Bits	3	4	5	6	7	8	9	10	11	12
1	17	39	39	27	43	41	23	39	19	35
3	51	11	11	47	1	5	59	11	57	33
9	25	33	33	13	3	15	49	33	43	27
15	1	55	55	21	5	25	61	55	29	45
21	27	51	51	55	7	35	17	51	61	63
33	49	7	7	5	11	55	45	7	41	43
63	47	25	25	37	21	23	49	25	55	11

*Frequency normalization factor = sampling frequency/128

Table Ib. Frequency of Peak Harmonic Distortion

Input Frequency* \ Total Noise Power (dB) \ No. of Bits	3	4	5	6	7	8	9	10	11	12
1	-20.99	-28.31	-34.47	-41.01	-47.30	-52.66	-60.32	-65.35	-71.35	-76.30
3	-20.99	-28.31	-34.47	-41.01	-47.30	-52.66	-60.32	-65.35	-71.35	-76.27
9	-20.99	-28.31	-34.47	-41.01	-47.30	-52.66	-60.31	-65.35	-71.32	-76.22
15	-20.99	-28.31	-34.47	-41.01	-47.30	-52.66	-60.33	-65.34	-71.35	-76.16
21	-20.99	-28.31	-34.47	-41.01	-47.29	-52.66	-60.30	-65.36	-71.36	-76.14
33	-20.99	-28.31	-34.47	-41.01	-47.29	-52.65	-60.31	-65.30	-71.35	-76.16
63	-20.99	-28.31	-34.47	-41.02	-47.29	-52.68	-60.31	-65.37	-71.42	-76.81

*Frequency normalization factor = sampling frequency/128

Table II. Total Quantization Noise Power

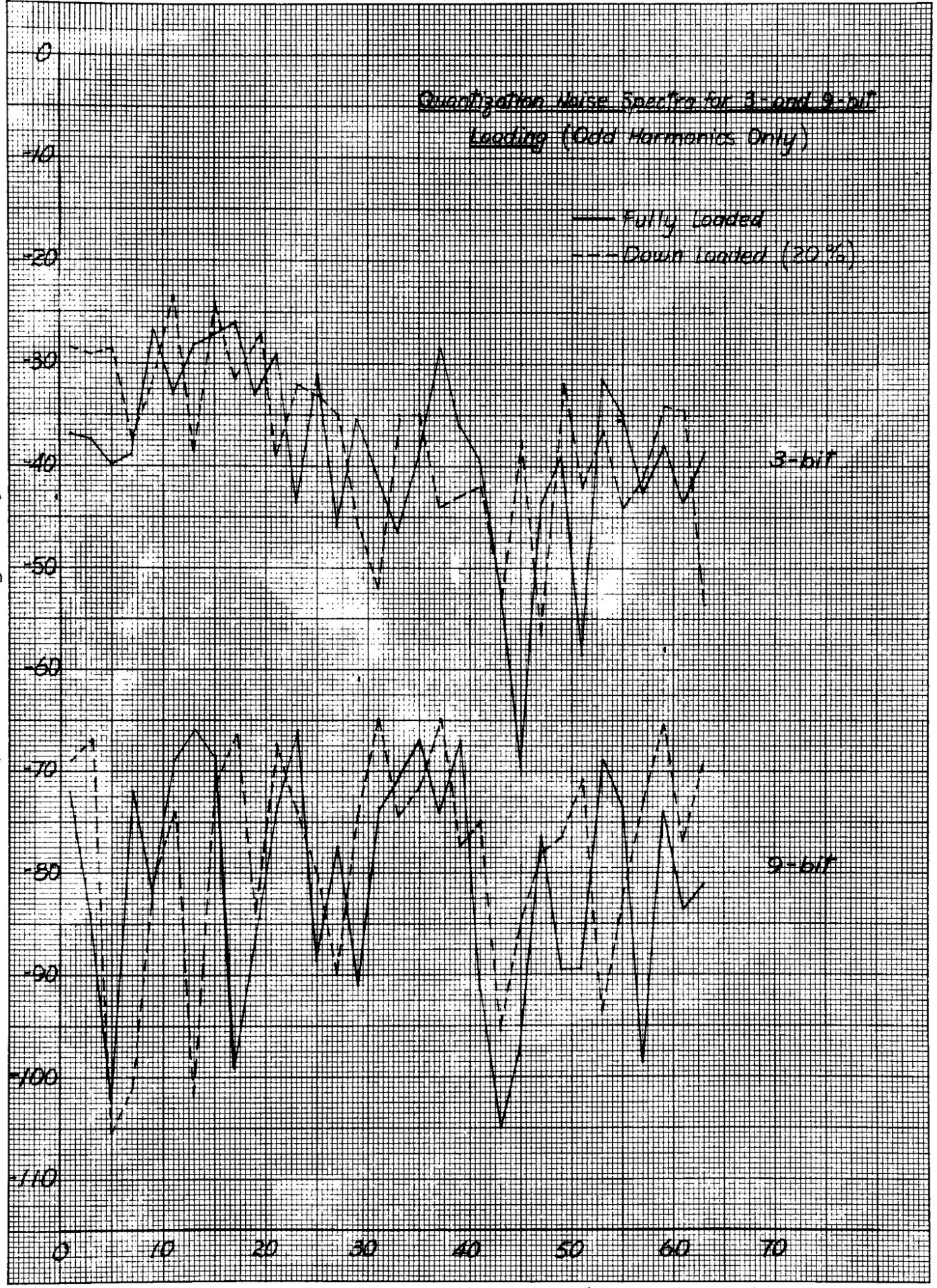

Quantization Noise Spectra for 3- and 9-bit Loading (Odd Harmonics Only)
Fully Loaded
Down Loaded (20%)
3-bit
9-bit
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
Spectral Density (dB)
0
10
20
30
40
50
60
70
Frequency (Sampling Rate/128)

Figure 3

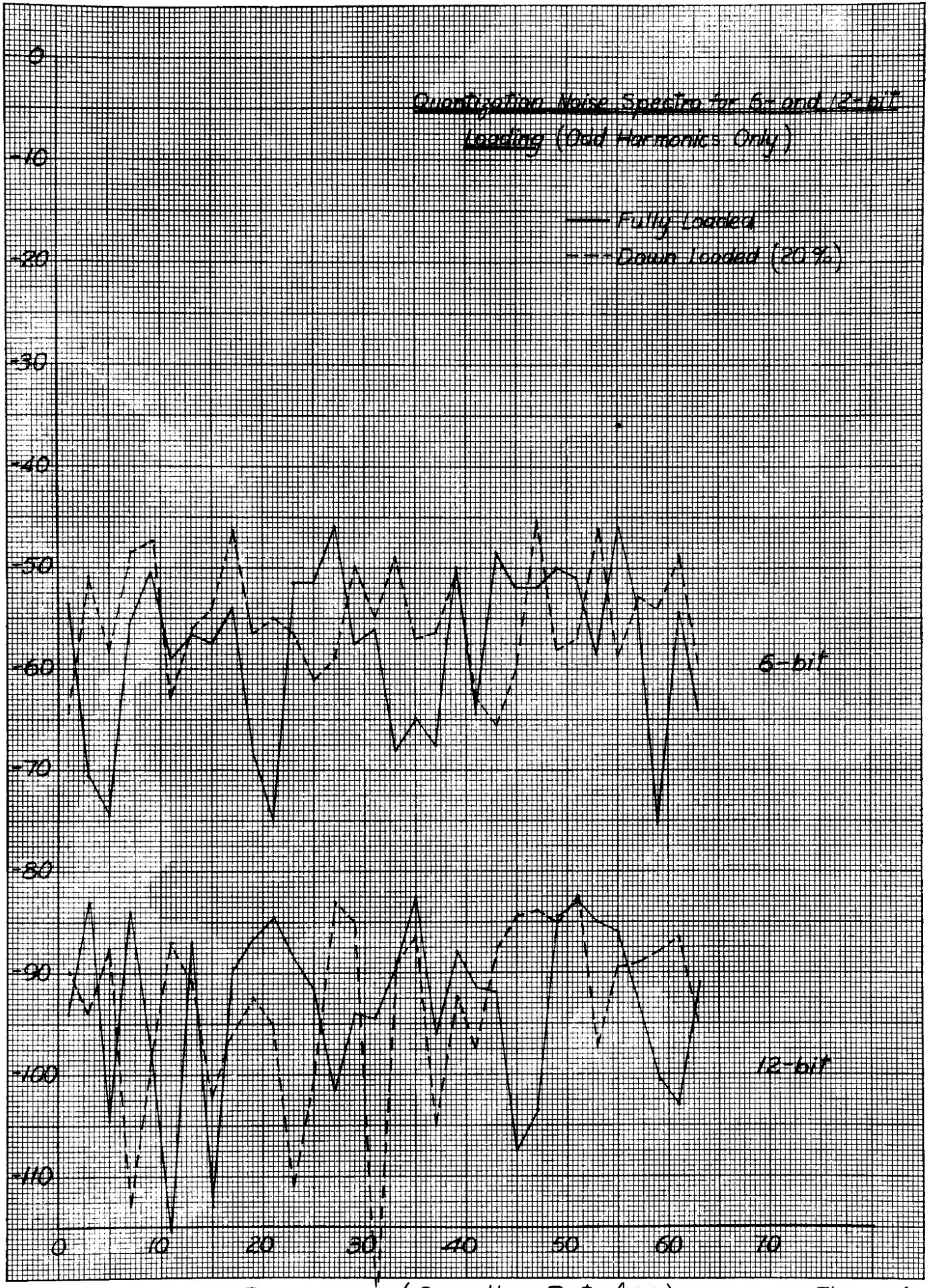

Quantization Noise Spectra for 6- and 12-bit Loading (Odd Harmonics Only)
Fully Loaded
Down Loaded (20%)
6-bit
12-bit
Spectral Density (dB)
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
0
10
20
30
40
50
60
70
Frequency (Sampling Rate/128)

Figure 4

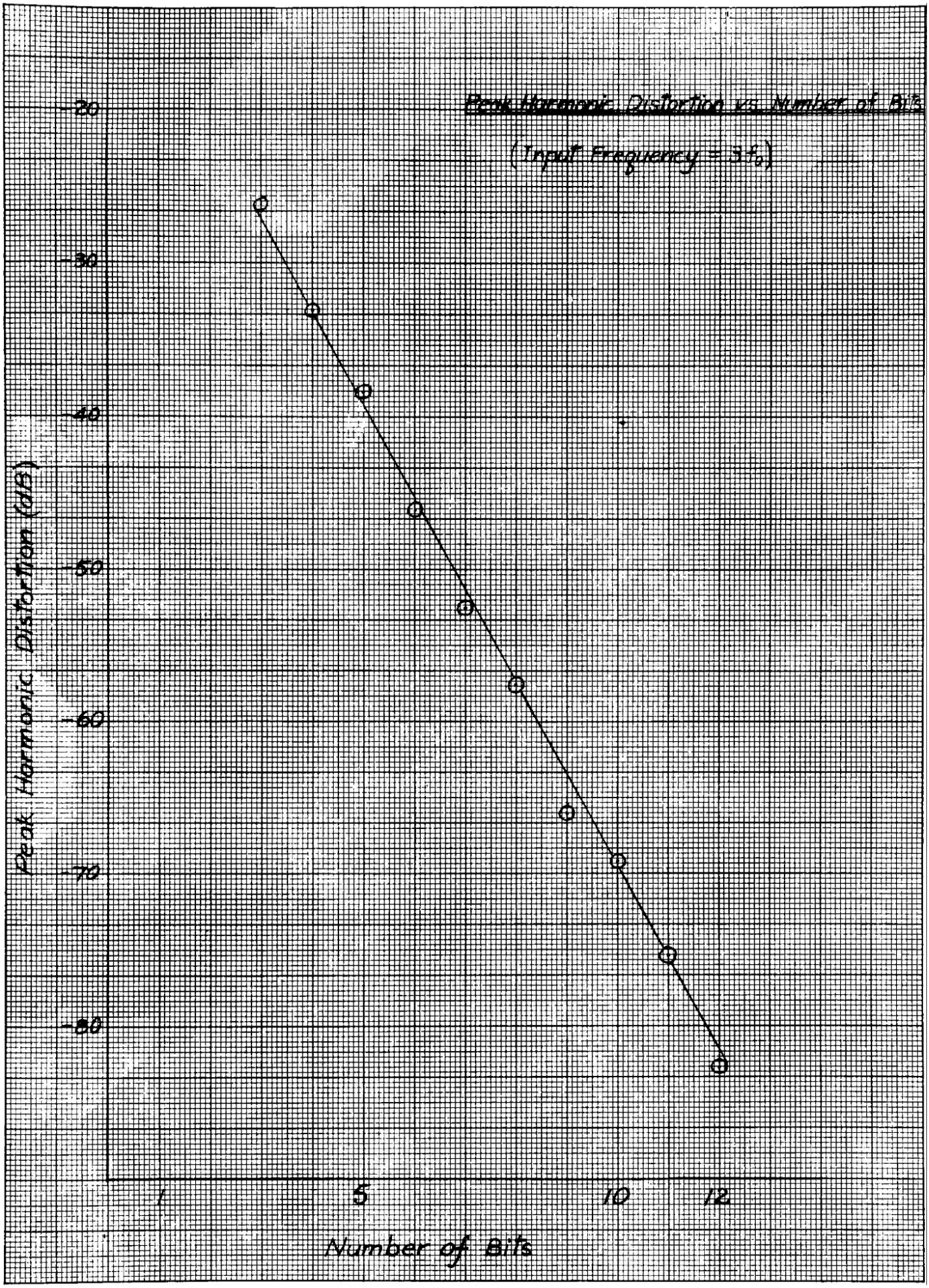

Peak Harmonic Distortion vs. Number of Bits
(Input Frequency = 3.f0)
-20
-30
-40
-50
-60
-70
-80
Peak Harmonic Distortion (dB)
1
5
10
12
Number of Bits

Figure 5

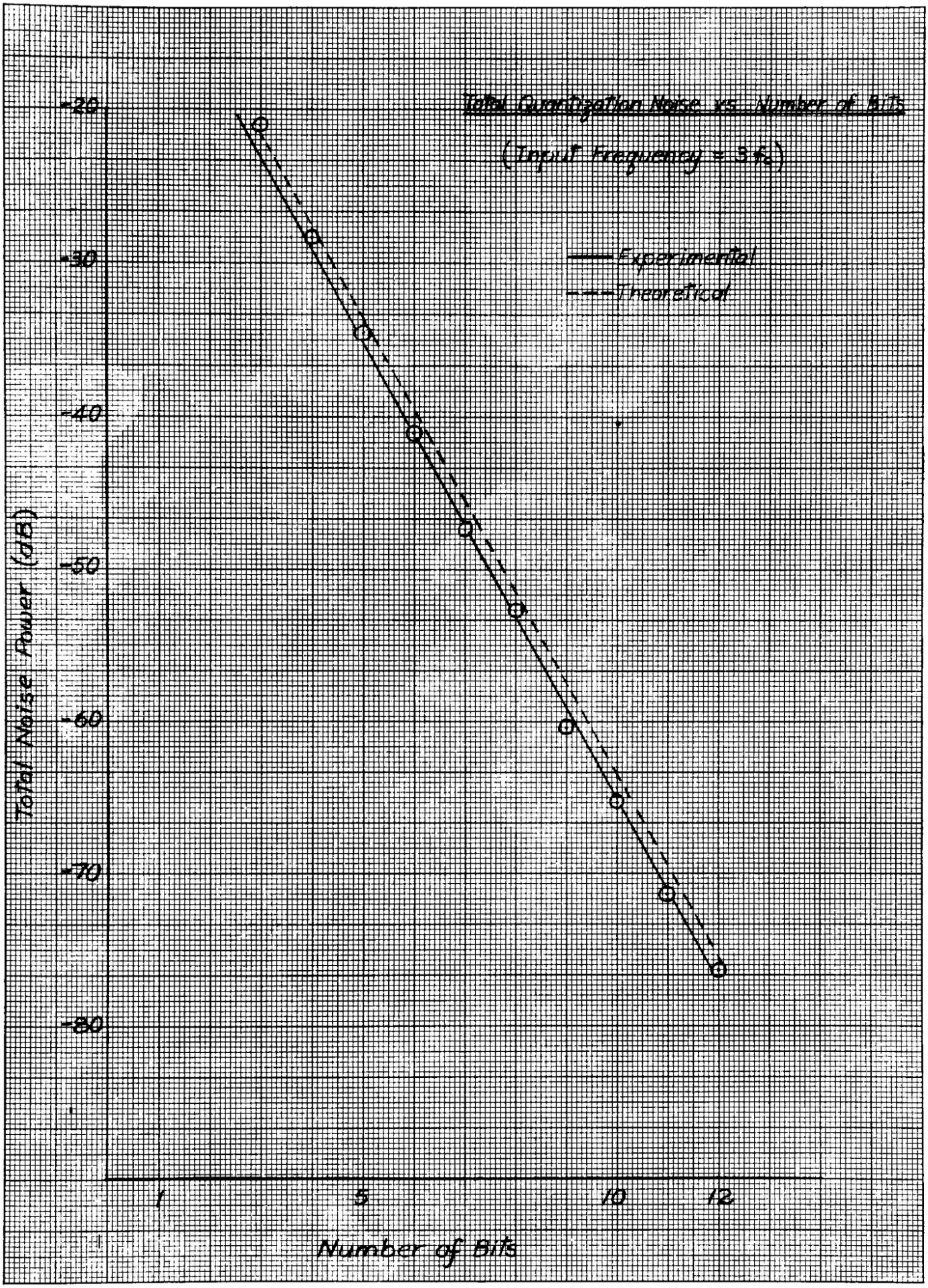

Total Quantization Noise vs. Number of Bits
(Input Frequency = 3 fs)
Experimental
Theoretical
-20
-30
-40
-50
-60
-70
-80
Total Noise Power (dB)
1
5
10
12
Number of Bits

Figure 6

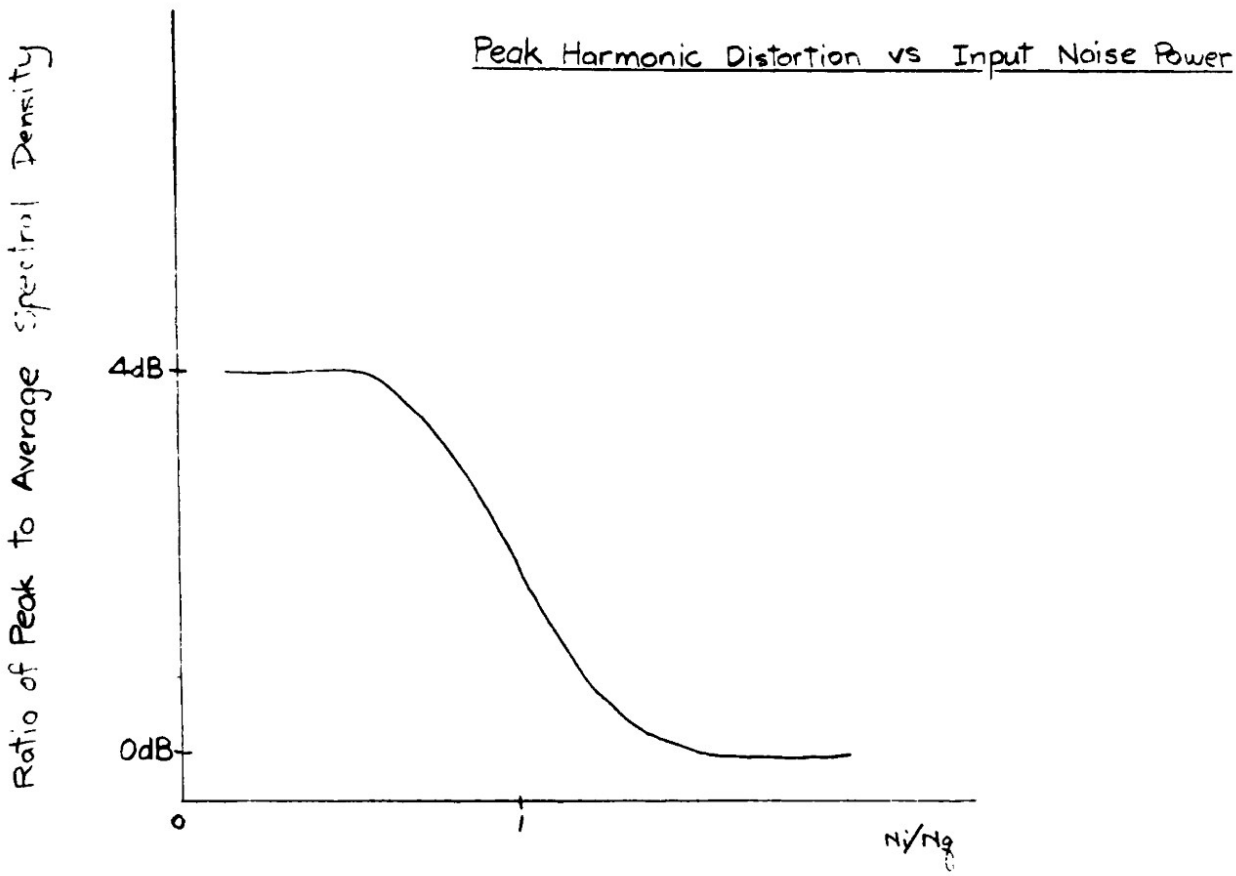

Peak Harmonic Distortion vs Input Noise Power
Ratio of Peak to Average Spectral Density
4dB
0dB
0
1
Ni/Ng

Figure 7.

V. CONCLUSION

In this thesis the quantization noise spectra for single frequency inputs were investigated in terms of the maximum spectral line. We found that the peak harmonic distortion follows a -6dB/bit curve and is about 4dB above the average noise spectral density, regardless of the number of bits in the quantizer and the sampling frequency.

Appendix A. APPROXIMATION TO GAUSSIAN NOISE

In this section we shall demonstrate how well a synthesized signal which consists of 63 statistically independent frequency components with uniformly random phases, approximates bandlimited white Gaussian noise in terms of the initial moments of the two random variables.

The synthesized signal can be written as:

$$x = \sum_{n=1}^{63} A \sin(n\omega_o t+\phi_n) = \sum_{n=1}^{63} x_n \qquad (A.1)$$

where A is a constant and ϕ_n's are statistically independent phases uniformly distributed between 0 and 2π. Since x is a finite sum of cosine waves with identical amplitude, the signal has a flat bandlimited spectrum.

The central moments of x can be computed from either the probability density function (pdf) of x or its characteristic function which is the Fourier transform of the pdf. However since x is the sum of 63 independent random variables, the characteristic function is simply the product of the individual characteristic functions. Therefore the central moments of x will be determined by sucessively differentiating the characteristic function.

The characteristic function of x_n is given by:

$$G_n(v) = \langle \exp[jvx_n] \rangle = \int_{-\infty}^{\infty} \exp[jvA\sin(n\omega_o t+\phi)]\, p(\phi)\, d\phi$$

substituting in $p(\phi) = 1/2\pi$ for $0 \leq \phi \leq \pi$, which is the pdf of ϕ.

$$G_n(v) = (1/2\pi) \int_{o}^{2\pi} \exp[jvA \sin(n\omega_o t+\phi)]\, d\phi$$

$$G_n(v) = (1/2\pi) \int_{n\omega_o t}^{n\omega_o t + 2\pi} \exp[jvA \sin(u)]\, du \qquad u = n\omega_o t + \phi$$

$$= J_o(vA) \qquad (A.2)$$

where $J_o(z)$ is the Bessel function of the first kind and zeroth order. Notice that equation (A.2) is independent of n and t. It follows that the characteristic function of x is given by

$$G(v) = \prod_{n=1}^{63} G_n(v)$$

$$= [J_o(vA)]^{63} \qquad (A.3)$$

and the k-th initial moment of x is

$$m_k = j^{-k} G^{(k)}(0) \qquad (A.4)$$

For our purposes, we shall evaluate m_k for k=1,2...6, and to avoid any confusion of the superscript for derivatives and powers we shall denote the n-th derivative of $J_o(vA)\Big|_{v=0}$ as J_n.

$$m_1 = -j\ 63\ J_o^{62} J_1$$

$$m_2 = -63(62 J_o^{61} J_1^2 + J_o^{62} J_2)$$

$$m_3 = j63(62 \cdot 61\ J_o^{60} J_1^3 + 62\ J_o^{61} \cdot 3\ J_1 J_2 + J_o^{62} J_3)$$

$$m_4 = 63(62\text{-}61 \cdot 60\ J_o^{59} J_1^4 + 62 \cdot 61\ J_o^{60} \cdot 6\ J_1^2 J_2 + 62\ J_o^{61} \cdot 3\ J_2^2 + 62\ J_o^{61} \cdot 4\ J_1 J_3 + J_o^{62} J_4)$$

$$m_5 = -j63(62\cdot61\cdot60\cdot59\ J_o^{58} J_1^5 + 62\cdot61\cdot60\ J_o^{59}\cdot10\ J_1^3 J_2 + 62\cdot61\ J_o^{60}(15\ J_1 J_2^2 + 10\ J_1^2 J_3) + 62\ J_o^{61}(10\ J_2 J_3 + 5\ J_1 J_4) + J_o^{62} J_5)$$

$$m = -63(62\cdot61\cdot60\cdot59\cdot58\ J_0^{57} J_1^6 + 62\cdot61\cdot60\cdot59\ J_o^{58} 15\ J_1^4 J_2 + 62\cdot61\cdot60\ J_o^{59} (45\ J_1^2 J_2^2 + 20\ J_1^3 J_3) + 62\cdot61\ J_o^{60}(15\ J_2^3 + 60\ J_1 J_2 J_3 + 15\ J_1^2 J_4) + 62\ J_o^{61}(10 J_3^2 + 15\ J_2 J_4 + 6\ J_1 J_5) + J_o^{62} J_6)$$

To determine the J_n's, the series expansion of $J_o(z)$ will be used.

$$J_o(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{2^{2n}(n!)^2}$$

$$= 1 - z^2/4 + z^4/64 - z^6/2304 + \ldots \qquad (A.5)$$

$$J_o = J_o(Av)\Big|_{v=0} = 1$$

$$J_1 = (-A^2 v/2 + A^4 v^3/16 - A^6 v^5/384 + \ldots)\Big|_{v=0} = 0$$

$$J_2 = (-A^2/2 + 3\ A^4 v^2/16 - 5\ A^6 v^4/384 + \ldots)\Big|_{v=0} = -A^2/2$$

$$J_3 = (3\ A^4 v/8 - 5\ A^6 v^3/96 + \ldots)\Big|_{v=0} = 0$$

$$J_4 = (3A^4/8 - 5\ A^6 v^2/32 + \ldots)\Big|_{v=0} = 3A^4/8$$

$$J_5 = (-5\ A^6 v/16 + \ldots)\Big|_{v=0} = 0$$

$$J_6 = (-5\ A^6/16 + \ldots)\Big|_{v=0} = -5A^6/16$$

Substituting the J_n's into the m_K's, we obtain

$$m_1 = 0$$

$$m_2 = 63\ A^2/2$$

$$m_3 = 0$$

$$m_4 = 3 \cdot 63^2 (A^2/2)^2 (1 - 1/126)$$

$$m_5 = 0$$

$$m_6 = 5 \cdot 3 \cdot 63\ (A^2/2)^3 (1 - 5/(6(63^2))$$

The central moments of the Gaussian distribution are given by

$$m'_{2k} = \sigma^{2k} (2k-1)!!, \quad m'_{2k+1} = 0 \qquad (A.6)$$

where σ^2 is the variance, and $k!! = 1 \cdot 3 \cdot 5 \cdots k$, for odd k.

Evauating equation (A.6) for k=1,2,3, we obtain

$$m'_1 = 0$$

$$m'_2 = \sigma^2$$

$$m'_3 = 0$$

$$m'_4 = 3 \cdot \sigma^4$$

$$m'_5 = 0$$

$$m'_6 = 5 \cdot 3\,\sigma^6$$

Since the variance of the sum of N independent random variables is the sum of the individual variances,

$$\sigma^2 = 63\ A^2/12.$$

Comparing the two sets, $\{m_k\}$ and $\{m'_k\}$, we find that there are discrepancies only for k=4 and 6, and these are 0.0079 and 0.0002 respectively.

Consequently we can conclude that the synthesized noise signal used for the computer simulation is a good approximation to band-limited white, Gaussian noise.

Appendix B. ERGODICITY OF SYNTHESIZED NOISE

In this section we shall demonstrate that the synthesized noise signal as given by equation (A.1) is ergodic only up to the second moment. The signal can be expressed as:

$$x(t) = A \sum_{n=1}^{63} \sin(n\omega_o t + \phi_n)$$

Six initial moments have been calculated in appendix A by ensemble averaging, and we shall repeat the calculation by time averaging.

$$m_1' = \overline{x(t)} = (1/T) \int_o^T x(t)\, dt \qquad \text{where } T = 2\pi/\omega_o$$

$$= (A/T) \sum_{n=1}^{63} \int_o^T \sin(n\omega_o t + \phi_n)\, dt$$

$$= 0 \qquad (B.1)$$

$$m_2' = \overline{x^2(t)} = (1/T) \int_o^T x^2(t)\, dt$$

$$= (A^2/\pi T) \sum_{m=1}^{63} \sum_{n=1}^{63} \int_o^T \sin(m\omega_o t + \phi_m) \sin(n\omega_o t + \phi_n)\, dt$$

$$= (A^2/2T) \sum_{m=1}^{63} \sum_{n=1}^{63} \int_o^T [\cos([m-n]\omega_o t + \phi_m - \phi_n) - \cos([m+n]\omega_o t + \phi_m + \phi_n)]\, dt$$

$$= (A^2/2T) \sum_{n=1}^{63} \int_o^T \cos([n-n]\omega_o t + \phi_n - \phi_n)\, dt - 0$$

$$= 63\ (A^2/2) \qquad (B.2)$$

$$m_3' = \overline{x^3(t)} = (1/T)\int_o^T x^3(t)\,dt$$

$$= (A^3/T)\int_o^T \sum_{l=1}^{63}\sum_{m=1}^{63}\sum_{n=1}^{63} \sin(l\omega_o t+\phi_l)\,\sin(m\omega_o t+\phi_m)\,\sin(n\omega_o t+\phi_n)\,dt$$

$$= (A^3/2T)\sum_{l=1}^{63}\sum_{m=1}^{63}\sum_{n=1}^{63}\int_o^T \sin(l\omega_o t+\phi_l)[\cos[m-n]\omega_o t+\phi_m-\phi_n) - \cos([m+n]\omega_o t+\phi_m+\phi_n]\,dt$$

$$= (A^3/4T)\sum_{l=1}^{63}\sum_{m=1}^{63}\sum_{n=1}^{63}\int_o^T [\sin([l+m-n]\omega_o t+\phi_l+\phi_m-\phi_n)$$
$$+ \sin([l-m+n]\omega_o t+\phi_l-\phi_m+\phi_n)$$
$$- \sin([l-m-n]\omega_o t+\phi_l-\phi_m-\phi_n)$$
$$- \sin([l+m+n]\omega_o t+\phi_l+\phi_m+\phi_n)]\,dt$$

$$= (A^3/4)[\sum_{l=2}^{63}\sum_{\substack{m\,n\\ m+n=l}} \sin(-\phi_l+\phi_m+\phi_n) + \sum_{m=2}^{63}\sum_{\substack{l,n\\ l+n=m}} \sin(\phi_l-\phi_m+\phi_n)$$
$$+ \sum_{n=2}^{63}\sum_{\substack{l,m\\ l+m=n}} \sin(\phi_l+\phi_m-\phi_n) - 0]$$

$$= (3A^3/4)\sum_{l=2}^{63}\sum_{\substack{m,n\\ m+n=l}} \sin(-\phi_l+\phi_m+\phi_n) \qquad (B.3)$$

Comparing the three "time" moments to those calculated in Appendix A, we see that the first two agree, but the third moment given by equation (B.3) is dependent on the specific set of phase angles and its value is certainly not zero in all cases. Similarly we can induce that all the higher "time" moments are functions of the phase angles and they do not necessarily agree with the ensemble moments. Consequently we can conclude that the synthesized noise signal used in this thesis is not ergodic for moments higher than the second, so the computer generated noise samples do not have consistent statistical properties.

REFERENCES

1. M. Schwartz, Information Transmission, Modulation, and Noise, McGraw-Hill, New York, 1959, Sec.6.7.
2. Special issue on Process Computers, Proc. IEEE, vol.58, Jan. 1970.
3. C.E. Shannon, "Communication in the Presence of Noise," Proc. IRE, vol.37, 1949, pp.10-21.
4. W. Bennett, "Spectra of Quantized Signals," Bell Sys. Tech. J., vol.27, July 1948, pp.446-472.
5. P.F. Panter, Modulation, Noise, and Spectral Analysis, McGraw-Hill, New York, 1965, pp.633-641.
6. W.B. Davenport and W.L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York, 1958, Chap.13.
7. E.D. Banta, "On the Autocorrelation Function of Quantized Signal Plus Noise," IEEE Trans. Inform. Theory, vol. IT-11, Jan.1965, pp.114-117.
8. H.E. Rowe, Signals and Noise in Communication System, Van Nostrand, New York, 1965, Sec. 4.0.
9. J. Max, "Quantizing for Minimum Distortion," IRE Trans. Inform. Theory, vol. IT-6, March 1960, pp.7-12.
10. V.D. Algazi, "Useful Approximations to Optimum Quantization," IEEE Trans. Comm. Technology, vol. COM-14, June 1966, pp.297-301.
11. W.J. Hurd, "Correlation Function of Quantized Sine Wave Plus Gaussian Noise," IEEE Trans. Inform. Theory, vol. IT-13, Jan.1967, pp.65-68.
12. A. Papoulis, "Error Analysis in Sampling Theory," Proc. IEEE, vol.54, July 1966, pp.947-955.