

Duplicate of 38413

N 7 2 30 19 1

MSC-07225

NASA TECHNICAL MEMORANDUM

NASA TM X-58097  
August 1972



CASE FILE  
COPY

# QUANTIZATION NOISE IN DIGITAL SPEECH

A Thesis Presented to the  
Faculty of the Graduate School  
of the University of Houston  
in Partial Fulfillment of the  
Requirements for the Degree of  
Master of Science in Electrical Engineering

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

MANNED SPACECRAFT CENTER

HOUSTON, TEXAS 77058

1. Report No.		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle  <b>QUANTIZATION NOISE IN DIGITAL SPEECH</b>				5. Report Date <b>August 1972</b>	
				6. Performing Organization Code	
7. Author(s) <b>Oron L. Schmidt, MSC</b>				8. Performing Organization Report No. <b>NASA TM X-58097</b>	
9. Performing Organization Name and Address  <b>Manned Spacecraft Center Houston, Texas 77058</b>				10. Work Unit No.  <b>914-50-00-00-72</b>	
				11. Contract or Grant No.	
				13. Type of Report and Period Covered <b>Technical Memorandum</b>	
12. Sponsoring Agency Name and Address  <b>National Aeronautics and Space Administration Washington, D.C. 20546</b>				14. Sponsoring Agency Code	
15. Supplementary Notes					
16. Abstract  <p>The amount of quantization noise generated in a digital-to-analog converter is dependent on the number of bits or quantization levels used to digitize the analog signal in the analog-to-digital converter. The minimum number of quantization levels and the minimum sample rate are derived for a digital voice channel. The minimum calculated parameters for a digital voice channel with 100 percent sentence intelligibility are 16 quantization levels and a sample rate of 5000 samples per second. Lowpass filters at the input of the analog-to-digital converter and the output of the digital-to-analog converter must have a 3 db cutoff frequency of 2000 Hz and a minimum rolloff of 36 db per octave. Laboratory results show that the calculated sample rate is slightly optimistic. A sample rate of 6000 samples per second and lowpass filters with a 3 db cutoff of 2400 Hz are required for 100 percent sentence intelligibility.</p> <p>Consonant sounds are the first speech components to be degraded by quantization noise. A compression amplifier can be used to increase the weighting of the consonant sound amplitudes in the analog-to-digital converter. An expansion network must be installed at the output of the digital-to-analog converter to restore the original weighting of the consonant sounds. This technique results in 100 percent sentence intelligibility for a sample rate of 5000 samples per second, eight quantization levels, and lowpass filters with a 3 db cutoff of 2000 Hz.</p>					
17. Key Words (Suggested by Author(s))  <ul style="list-style-type: none"> <li>* Quantization Errors</li> <li>* Masking of Speech</li> </ul>			18. Distribution Statement		
19. Security Classif. (of this report)  <b>None</b>		20. Security Classif. (of this page)  <b>None</b>		21. No. of Pages  <b>66</b>	
				22. Price*	

NASA TM X-58097

**QUANTIZATION NOISE IN DIGITAL SPEECH**

**Oron L. Schmidt  
Manned Spacecraft Center  
Houston, Texas 77058**

QUANTIZATION NOISE IN  
DIGITAL SPEECH

---

A Thesis

Presented to  
the Faculty of the Department of Electrical Engineering  
University of Houston


---

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science

---


by  
Oron L. Schmidt  
December 1970

QUANTIZATION NOISE IN  
DIGITAL SPEECH

  
Oron L. Schmidt

Approved:

Director and Chairman  
of the Committee:

  
H. S. Hayre

Associate Advisor:

  
N. M. Shehadeh

Associate Advisor:

  
R. L. Motard

Dean, College of  
Engineering

  
C. V. Kirkpatrick

## ACKNOWLEDGEMENTS

The author wishes to acknowledge the guidance and counsel received from Dr. H. S. Hayre, faculty advisor, and the members of the thesis committee.

## TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION . . . . .	1
II. WAVEFORM CHARACTERISTICS OF SPEECH . . . . .	4
III. SPEECH DIGITIZATION . . . . .	8
IV. QUANTIZATION NOISE . . . . .	13
V. MINIMIZING QUANTIZATION NOISE . . . . .	47
VI. CONCLUSIONS AND RECOMMENDATIONS . . . . .	53
BIBLIOGRAPHY . . . . .	55
APPENDIX A . . . . .	56

# LIST OF FIGURES

FIGURE		PAGE
2-1.	Vowel Sound Spectra . . . . .	6
2-2.	Consonant Sound Spectra . . . . .	7
3-1.	Sampled Speech Waveform . . . . .	11
3-2.	Quantized Speech Waveform . . . . .	12
4-1.	Quantization Error . . . . .	14
4-2.	Maximum Output $S_{rms}/N_{rms}$ Versus the Number of Quantization Levels . . . . .	17
4-3.	Input S/N Versus Output S/N . . . . .	18
4-4.	500 Hz Sinewave Sampled at 2000 Hz and Quantized with 16 Levels . . . . .	19
4-5.	2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 32 Levels . . . . .	21
4-6.	2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 16 Levels . . . . .	22
4-7.	2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 8 Levels . . . . .	23
4-8.	2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 4 Levels . . . . .	24
4-9.	ADC Sampling Function and DAC Approximating Function . . . . .	25
4-10.	Sampling Function Synchronous and Asynchronous with ADC Input . . . . .	29



FIGURE	PAGE
4-11. Oscilloscope Waveform at Sampling Function Asynchronous with ADC Input (DAC Output) . . .	31
4-12. PSD of DAC Output for Speech Sampled at 4000 Samples per Second . . . . .	32
4-13. PSD of DAC Output for Speech Sampled at 2500 Samples per Second . . . . .	34
4-14. Optimizing the Lowpass Filter Cutoff Frequency	36
4-15. Sample Rate Versus W.I. for Speech Quantized with 8, 16, 32 and 64 Levels . . . . .	40
4-16. PSD of ADC Input . . . . .	42
4-17. PSD of DAC Output for Speech Quantized with 32 Levels . . . . .	43
4-18. PSD of DAC Output for Speech Quantized with 16 Levels . . . . .	44
4-19. PSD of DAC Output for Speech Quantized with 8 Levels . . . . .	45
4-20. PSD of DAC Output for Speech Quantized with 4 Levels . . . . .	46
5-1. Analog Nonlinear Encoding . . . . .	48
5-2. Sample Rate Versus WI for Companded Speech Quantized at 4, 8, 16 and 32 Levels . . . . .	50

## CHAPTER I

### INTRODUCTION

Digitizing speech for transmission over a communications link between a manned spacecraft and a ground station is required if the voice channel is to be integrated into a serial bit stream along with PCM (Pulse Code Modulation) telemetry data. If standard PCM techniques are used for digitizing the speech, the resulting bit rates can be high (about 40,000 bits per second). When the bandwidth of a communication link must be minimized to stay within the available transmitter power allotment, steps must be taken to keep this bit rate as low as possible and still meet the minimum channel quality requirements. The two parameters that determine the serial bit rate, namely, the sample rate and the number of quantization levels, can be reduced until sampling errors and quantization errors generate distortion products (noise) in sufficient quantity to render the decoded speech unacceptable.

Linear digital encoding is used by the Bell Telephone Company's digital voice communication links (Kleiner, 1969). The analog speech is sampled at 8000 samples per second. Quantization noise is kept very low by quantizing with 32 levels (5 bits per sample). This results in a very good communication link but requires a wide transmission bandwidth. Many of the telephone channels are digitized in Japan. Nonlinear

encoding techniques are used to keep the overall channel voice quality high with low bit rates (Taki, 1966).

The purpose of this thesis is to analyze the cause and develop techniques to minimize the effects of quantization noise in a digitized speech channel. This is accomplished in five chapters covering the waveform characteristics of speech, the process of digitizing speech, the sources of quantization noise, the causes of speech degradation by quantization noise, and the techniques to minimize quantization noise.

The characteristics of speech are analyzed in Chapter II. Human utterances can be divided into vowel and consonant sounds. Vowel sounds contain the most power and are formed by vibrating the vocal cords and manipulating an open vocal tract. Vowel sounds are essentially periodically damped sine waves. Consonant sounds are formed by constrictions in the vocal tract. Consonant waveforms are very complex and many times look like white noise in amplitude-versus-time plots.

Analog-to-digital conversion using PCM techniques is explained in Chapter III by using simple waveforms such as sine waves. The same techniques are then applied for digitizing speech. The speech waveforms are used instead of sine waves to show how speech is digitized to achieve the maximum intelligibility for a minimum resulting bit rate.

Optimum bandpass filters, sample rates, and quantization levels will be derived and discussed.

The optimum digital channel can be developed by minimizing the sample rate and the number of quantization levels until sampling errors and quantization noise degrade the channel performance below the minimum acceptable level. The source of sampling errors and quantization noise is analyzed in Chapter IV. The theoretical amount of quantization noise for several quantization levels is compared to laboratory experimental results. Optimum theoretical parameters for a low-bit-rate digital voice channel are also derived and compared with experimental results.

Techniques for minimizing the effects of quantization noise will be discussed in Chapter V. One of the successful methods of reducing the amount of quantization noise at low bit rates is to employ nonlinear quantization techniques. Greater weighting is given to the high frequency, low amplitude speech components. This is achieved through logarithmic compression prior to analog-to-digital conversion or a nonlinear ladder network is used in the analog-to-digital converter. The converse processing is carried out during or following the digital-to-analog conversion process to restore the original amplitude versus frequency relationships of the input speech. The results of laboratory tests are presented and discussed for an analog compression and expansion system that was used with a linear PCM system for speech.

## CHAPTER II

### WAVEFORM CHARACTERISTICS OF SPEECH

Spoken words and sentences are a sequence of sounds with little meaning individually, however, when one of these sounds replaces another in an utterance, the meaning is changed. These basic sounds are called phonemes. Human speech is a moderately complicated single (amplitude) function of time composed of smooth changes from one fairly stable sound (phoneme) to another. Each phoneme is produced by modulating either the radiated sound from the vocal cords (voiced sounds) or the sound produced by air forced through a constriction formed by the tongue, teeth, or lips (unvoiced sounds).

The vowel sounds or voiced sounds are produced by the vocal cord excitation of the vocal tract. The vocal tract, consisting of the lungs, trachea, larynx, mouth, and nasal passages, maintains a relatively stable configuration for the 12 vowel sounds in the English language. The position of the tongue with respect to the roof of the mouth changes the resonant frequency of the tract and consequently the harmonic components of the voiced sound. Each voiced sound corresponds to its own combination of harmonics or "formants." The fundamental harmonic is the pitch frequency or rate of vibration of the vocal cords. The pitch frequency normally

lies between 80 and 200 Hz for men and 150 to 350 Hz for women (Flanagan, 1965). The position of the tongue in the vocal tract determines which harmonics of the pitch frequency are reinforced or suppressed. Each vowel sound is characterized by its combination of harmonics (formants). Figure 2-1 shows the amplitude versus frequency distribution of eight of the twelve vowel sounds. Note that each sound can be characterized by four or less harmonics (formants) of the pitch frequency. The major power in many vowel sounds is centered around the third formant (400 - 800 Hz). More than eightypercent of the power in the vowel sounds is concentrated below 1500 Hz (Fletcher, 1961).

The consonant sounds are much more complex than the vowel sounds. These sounds are formed either by constrictions in the vocal tract which cause air turbulence (unvoiced) or by a combination of constrictions and vocal tract excitation (voiced). The intelligence bearing portion of consonant sounds is contained in the frequency range of 300 to 5000 Hz. The amplitude versus frequency spectrums of some consonant sounds, shown in Fig. 2-2, indicate that the major amplitude concentrations are between 2000 and 3000 Hz.

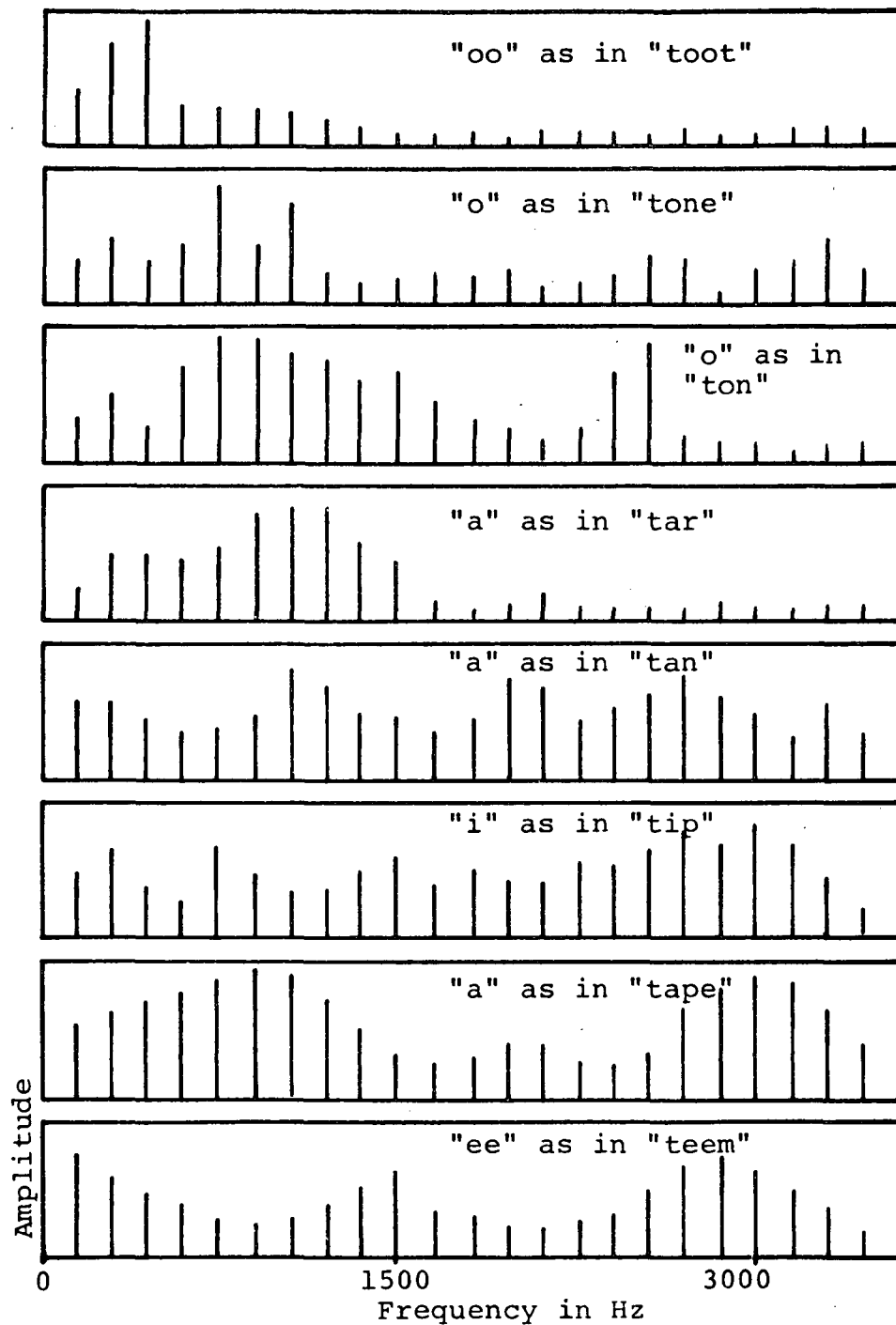


Fig. 2-1. Vowel Sound Spectra

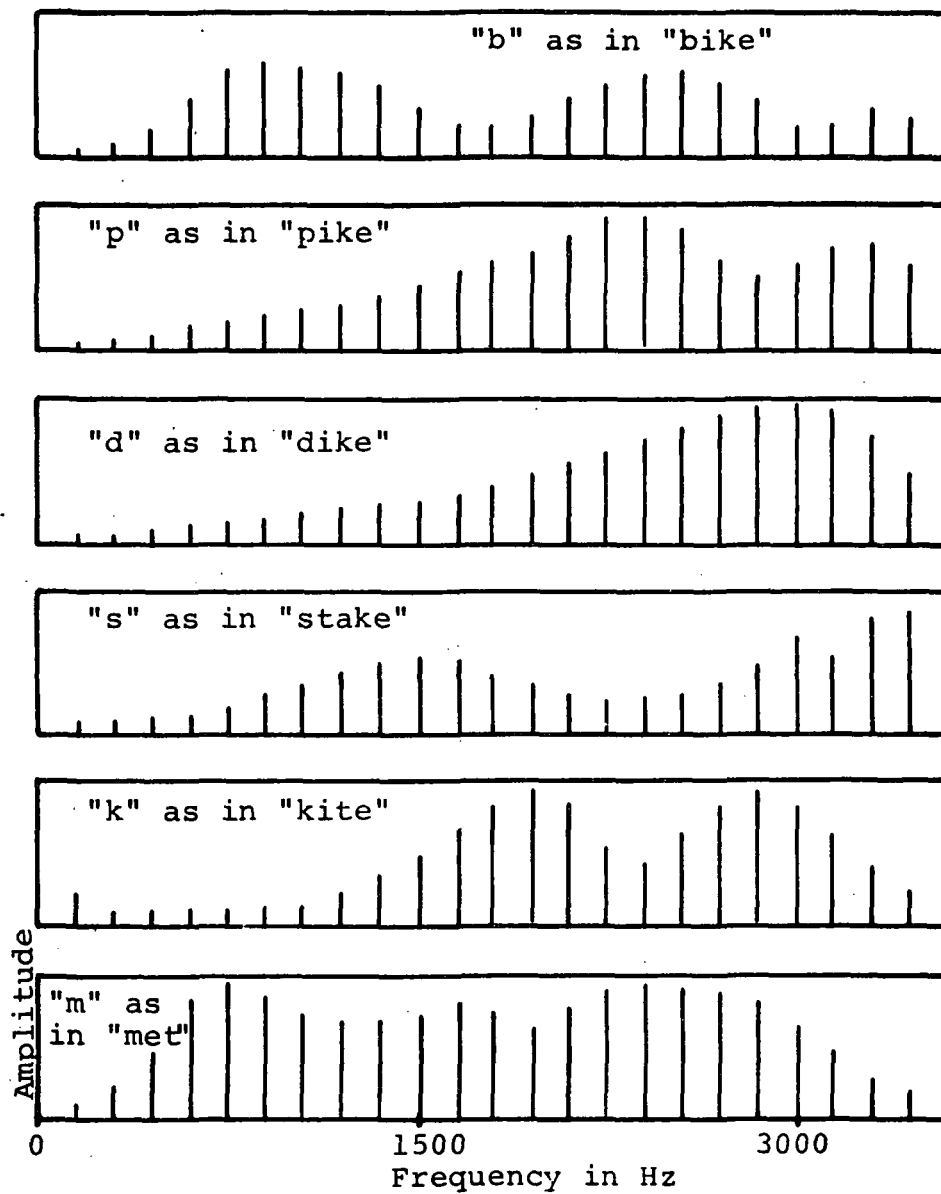


Fig. 2-2. Consonant Sound Spectra



## CHAPTER III

### SPEECH DIGITIZATION

A voice channel is degraded when the minimum output S/N requirements are not met. One solution of this problem is to encode the speech amplitude information using PCM encoding techniques. Encoded speech in the form of a binary bit stream can be time-division multiplexed with other digitized data bit streams. During the transmission, or following detection, the encoded bit stream can also be regenerated to preserve the original signal-to-noise relationship.

One of the first steps in developing a digital voice communication system is to define the voice channel performance requirements. If a voice channel is to provide sufficient quality to allow the perception of speaker identity and emotional status in addition to all utterances, the channel must have a wide frequency response (100 to 10,000 Hz) and a high speech-to-noise power ratio (in excess of 30 db) to reproduce all of the waveform components of vowel and consonant sounds. However, if the correct perception of all spoken words is the prime requirement, the required channel bandwidth can be much less. The listener can correctly interpret all sentence information, based on sentence and phonetic context, without using all of the higher frequency consonant components. Perception of key words in

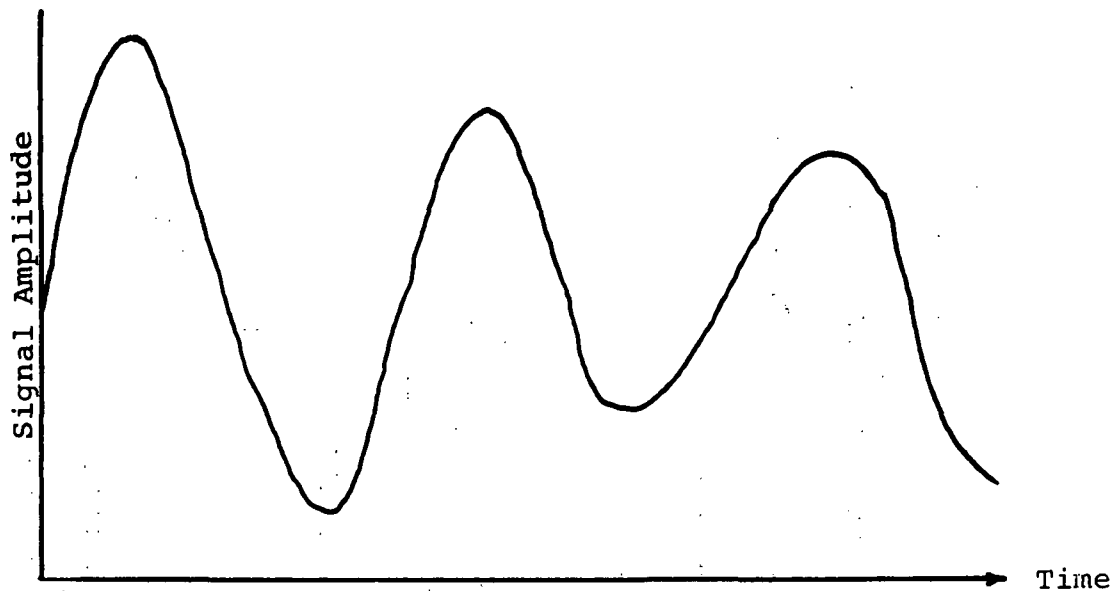
sentences defines the complete message since some words may be redundant or assumed by context. Perception of key harmonics of vowel and consonant sounds are sufficient to recognize the complete syllable or monosyllable word. Laboratory tests have shown that no degradation of channel word intelligibility results if speech is passed through a low-pass filter with a 3 db cutoff of 2.5 KHz and frequency rolloff of 36 db per octave (Schmidt, 1968). A word intelligibility degradation of only 5 percent was measurable when the 3 db cutoff was reduced to 1.5 KHz. Since the objective of this work is to develop an intelligible digital voice channel with a low bit rate, it is assumed that a minimum bandwidth analog speech spectrum is desired.

The first step in encoding an analog voice signal is to periodically sample it. Shannon's sampling theorem states that the minimum number of samples required to completely represent a time function  $f(t)$ , not containing frequency components higher than  $W$  Hz, can be expressed as (Taki, 1966)

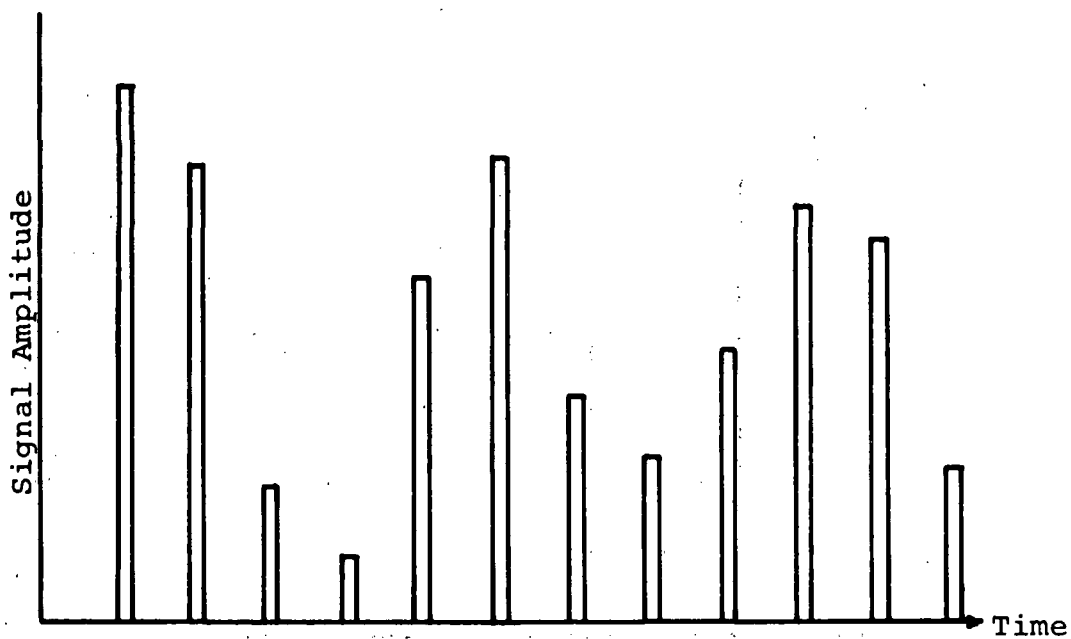
$$f(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2w}\right) \frac{\sin \pi(2wt-n)}{\pi(2wt-n)} \quad (3-1)$$

where  $f\left(\frac{n}{2w}\right)$  denotes the values of  $f(t)$  at time intervals spaced  $1/2 w$  seconds apart. The function  $f(t)$  is uniquely determined by this set of sampled values. For instance, if

a speech signal is filtered such that no components greater than 2500 Hz are present, the minimum sample rate would be 5000 samples per second. In a conventional analog-to-digital converter, the periodic samples of the input signal (Fig. 3-1(a)) form a PAM (Pulse Amplitude Modulated) waveform (Fig. 3-1(b)). To convert the PAM signal into a PCM signal, the pulse amplitude are binary encoded. The maximum input voltage range  $A$  is divided into equal discrete amplitude levels  $b$  volts apart. This is called the quantization process. The selection of the number of quantization levels determines how accurately the sample amplitudes are encoded. For illustration purposes it is assumed that eight levels are adequate for speech. A function representing the quantized waveform is shown in Fig. 3-2(b). Note that all voltages ranging between  $\pm b/2$  volts of a particular level are referred to that level. This approximation of the input signal by rounding off may result in a decoded signal that differs from the input signal. This quantization error is evident as noise in the decoder output. Each sample amplitude is represented by a binary word containing  $n$  bits for the  $2^n$  quantization levels used in the analog-to-digital conversion process. The binary words are then transmitted serially at a bit rate equal to the sample rate times the number of bits in each word. When the encoded bit stream reaches its destination, it is converted back to an analog speech waveform by a DAC (digital-to-analog converter).

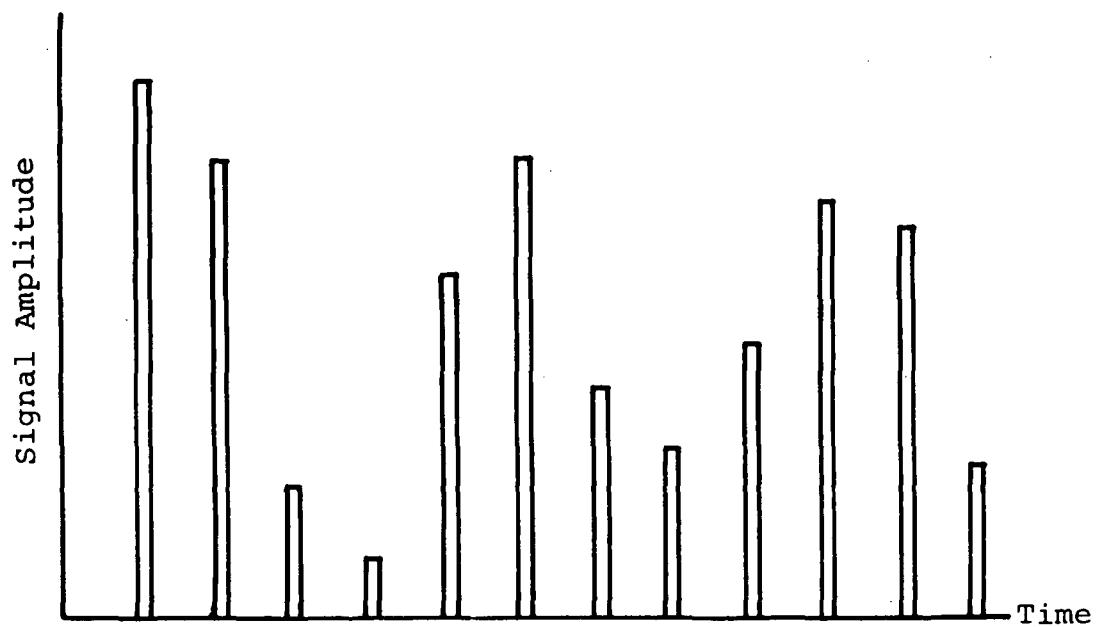


(a) Input Signal

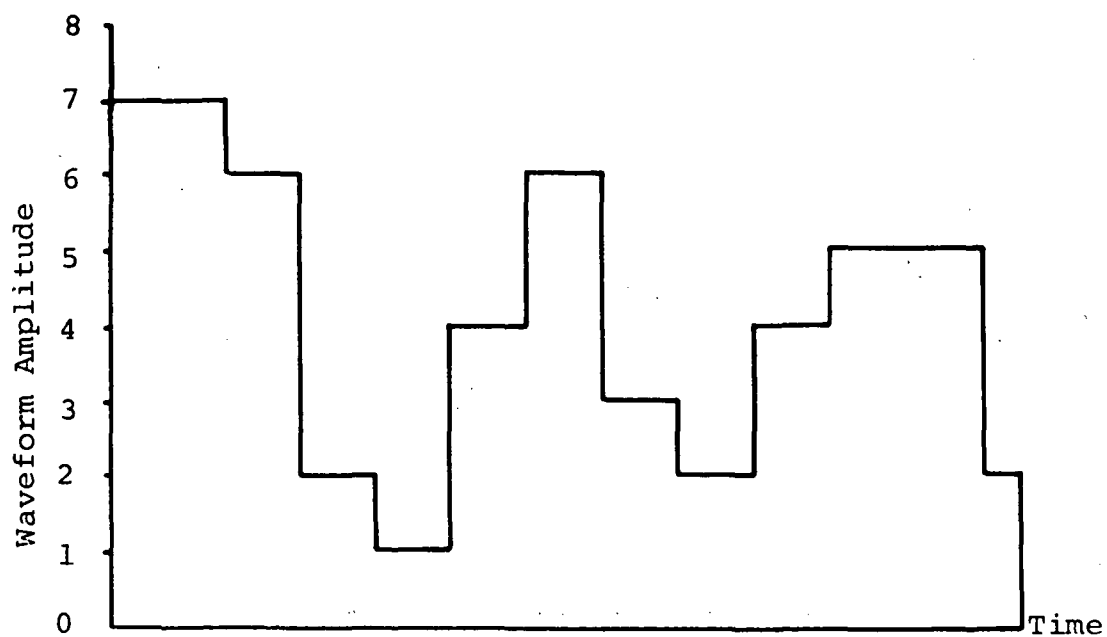


(b) PAM Output

Fig. 3-1. Sampled Speech Waveform



(a) PAM Input



(b) Quantized Waveform

Fig. 3-2. Quantized Speech Waveform

## CHAPTER IV

### QUANTIZATION NOISE

The minimum acceptable quantization noise at the digital voice channel's decoder output determines the minimum number of quantization levels required to encode each amplitude sample. To determine the amount of quantization noise to be expected, an expression for the output speech-to-noise ratio as a function of the number of quantization levels is developed.

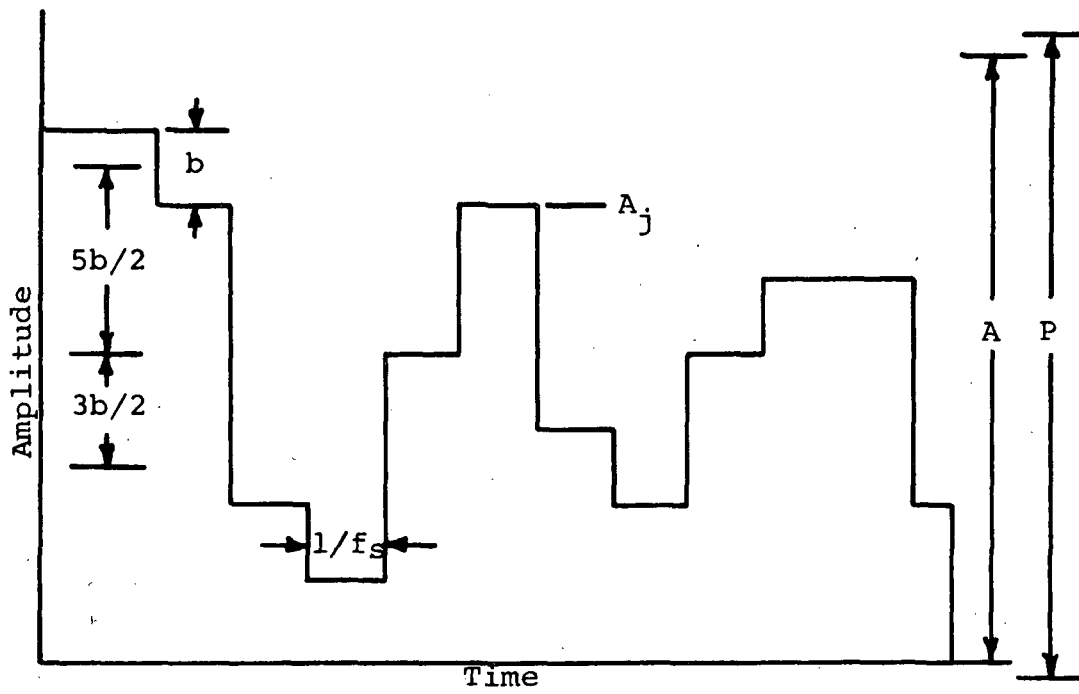
Let the signal into the ADC (analog-to-digital converter) be initially quantized in  $s$  levels, with  $b$  equal spacings between adjacent levels (see Fig. 4-1(a)). It is assumed that the input signal is referenced to zero volts (no dc component). The maximum voltage excursion  $P$  is divided up into  $b$  equal spacings such that

$$b = \frac{P}{s} \quad (4-1)$$

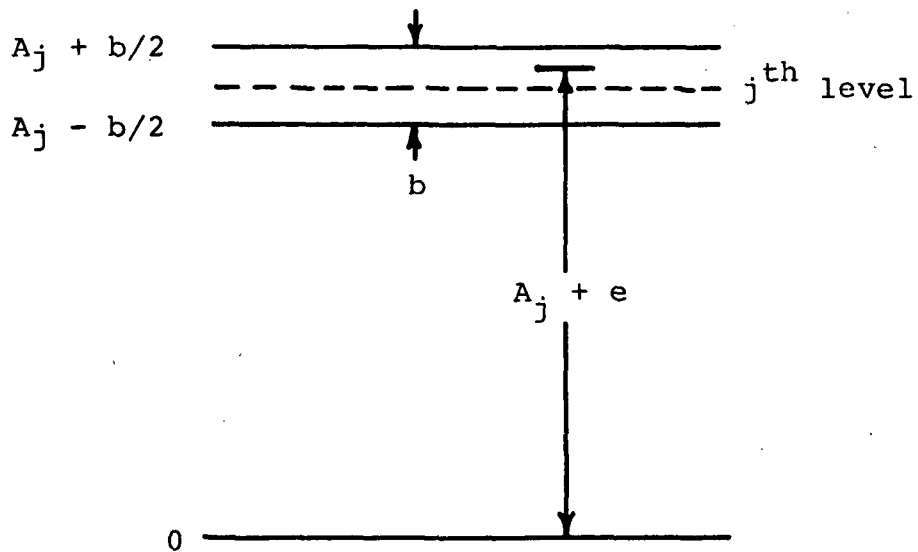
The quantized amplitudes are  $\pm b/2, \pm 2b/2, \pm 3b/2, \dots, \pm (s-1)(b/2)$ , and the quantized samples cover a range

$$A = (s-1)b \text{ volts} \quad (4-2)$$

In Chapter III it was stated that the quantization process introduces an uncertainty error since a sample appearing at



(a) Quantized ADC Input



(b) Region of Uncertainty

Fig. 4-1. Quantization Error

the DAC output could have been due to any input signal in the range  $A_j - b/2$  to  $A_j + b/2$  volts. To calculate the mean-squared error voltage, it is assumed that over a long period of time all voltage values in the region of uncertainty eventually appear the same number of times. The instantaneous signal amplitude is  $A_j + e$ , with  $-b/2 \leq e \leq b/2$ , where  $e$  represents the error voltage between the instantaneous (actual) signal and its quantized equivalent (Schwartz, 1959). For all values of  $e$  equally likely, the mean-squared value of  $e$  is

$$\overline{e^2} = 1/b \int_{-b/2}^{b/2} e^2 de = b^2/12 \quad (4-3)$$

The average error is zero and the rms error or output noise is  $b/\sqrt{12} = b/(2\sqrt{3})$  volts. The maximum DAC output signal-to-noise ratio (assuming an infinite S/N ratio at the input of the ADC) in terms of peak signal voltage to rms noise voltage is given by

$$\frac{S_{opv}}{N_{ovr}} = \frac{P}{b/2\sqrt{3}} = 2\sqrt{3} \text{ s} \quad (4-4)$$

the corresponding power ratio is

$$\frac{S_{op}}{N_{or}} = 12 \text{ s}^2 \quad (4-5)$$



or in decibels

$$\frac{S_{op}}{N_{or}} \text{ db} = 10.8 + 20 \log_{10} s \quad (4-6)$$

Peak speech power is approximately 14 db greater than rms speech power (Fletcher, 1961). Therefore the expression for maximum rms speech to rms noise power ratio can be expressed as

$$\frac{S_{or}}{N_{or}} \text{ db} = 20 \log_{10} s - 3.2 \quad (4-7)$$

Therefore the power ratio increases as the square of the number of quantization levels. The maximum output S/N ratio (rms to rms) for several different quantization levels is given in Fig. 4-2.

The laboratory experimental data (Fig. 4-3) shows that Eq. (4-7) is valid for obtaining an approximate maximum DAC output S/N ratio (Culver, 1969). It should also be noted that the output S/N is always less than the input S/N.

The power spectral distribution of quantization noise depends on the number of quantization levels and interaction between the input signal and the sampling function. Figure 4-4 shows the PSD (power spectral distribution) of the output of a DAC for a 500 Hz sinewave that was sampled at 2000 Hz and quantized using 16 levels. The harmonics shown occur at

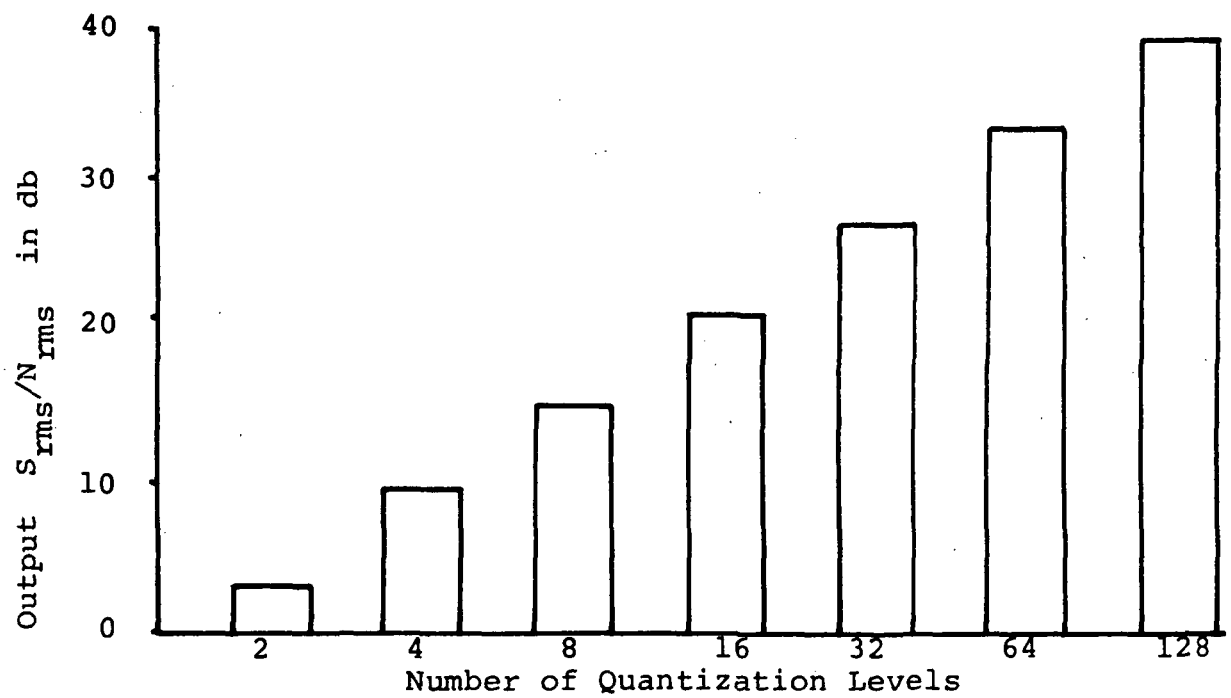


Fig. 4-2. Maximum Output  $S_{rms}/N_{rms}$  Versus the Number of Quantization Levels

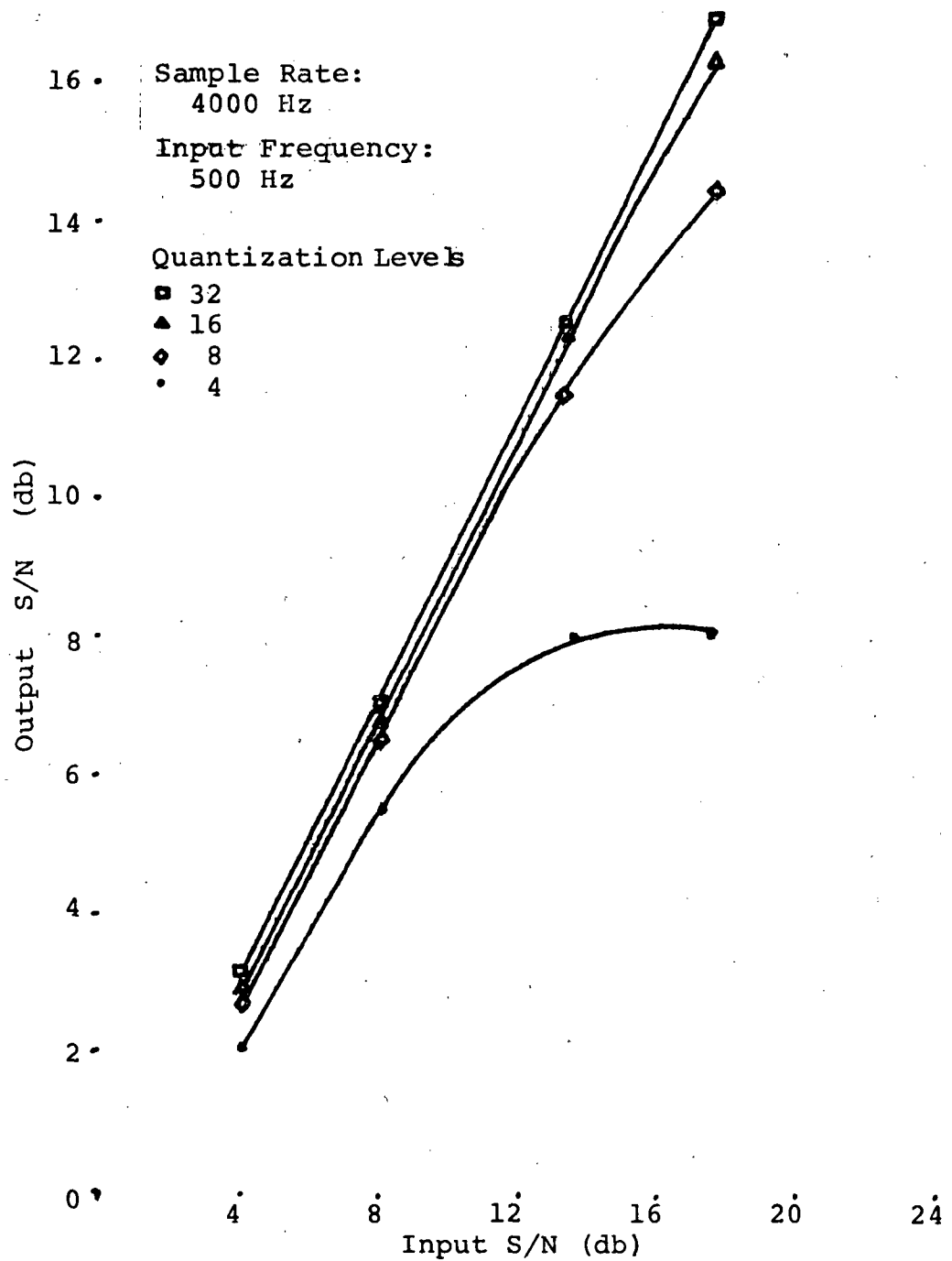


Fig. 4-3. Input S/N Versus Output S/N

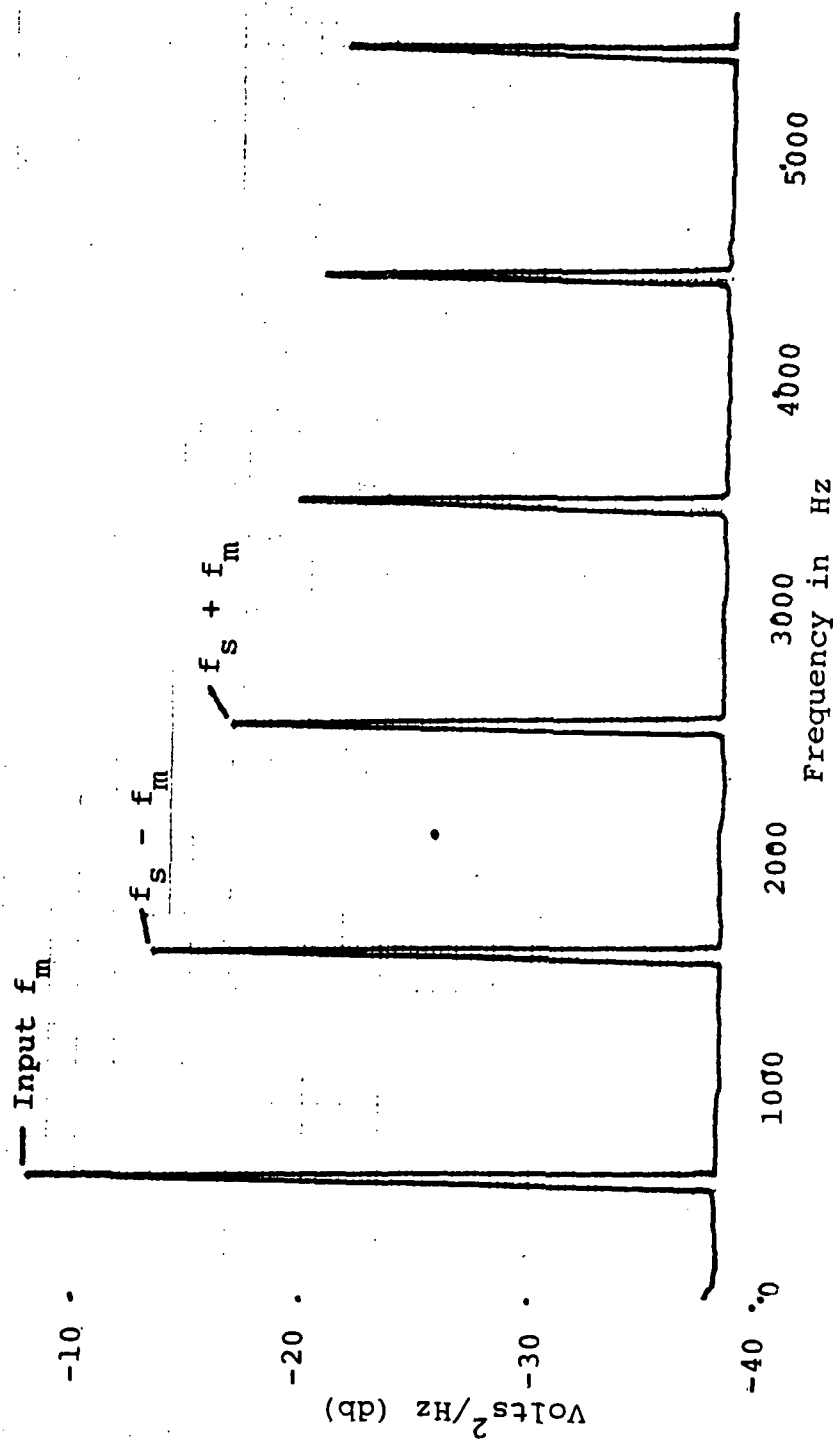


Fig. 4-4. 500 Hz Sine wave Sampled at 2000 Hz and Quantized with 16 Levels

$f_s \pm f_m$ ,  $2f_s \pm f_m$  and  $3f_s - f_m$  where  $f_s$  is the sampling frequency and  $f_m$  is the frequency of the input signal.

To investigate this phenomenon in more detail, a series of PSD's were made for an input signal of 2000 Hz sampled at 4150 Hz and quantized using 32, 16, 8 and 4 levels (see Figs. 4-5, 4-6, 4-7 and 4-8). The 4150 Hz sample rate was chosen to insure that the sum and difference harmonics would not be confused with the even-ordered harmonics of  $f_m$  or  $f_s$ . For 32 and 16 quantization levels, the only significant components were at  $f_m$ ,  $f_s \pm f_m$ , and  $2f_s - f_m$ . However, when 8 and 4 quantization levels were used, many more harmonics became significant.

The origin of all of the major harmonics appearing at the output of the DAC can be explained using Fourier analysis. The waveforms in the ADC and DAC can be represented by their Fourier series

$$f(t) = \frac{1}{T} \sum_{h=-\infty}^{\infty} C_h e^{jh\omega_0 t} \quad (4-8)$$

$$C_h = \int_{-T/2}^{T/2} f(t) e^{-jh\omega_0 t} dt \quad (4-9)$$

The sampling function  $S(t)$  is a periodic pulse waveform with a period  $T$  and a pulse width  $\Delta t$  (see Fig. 4-9(a)). The magnitude of the harmonic components of the sampling

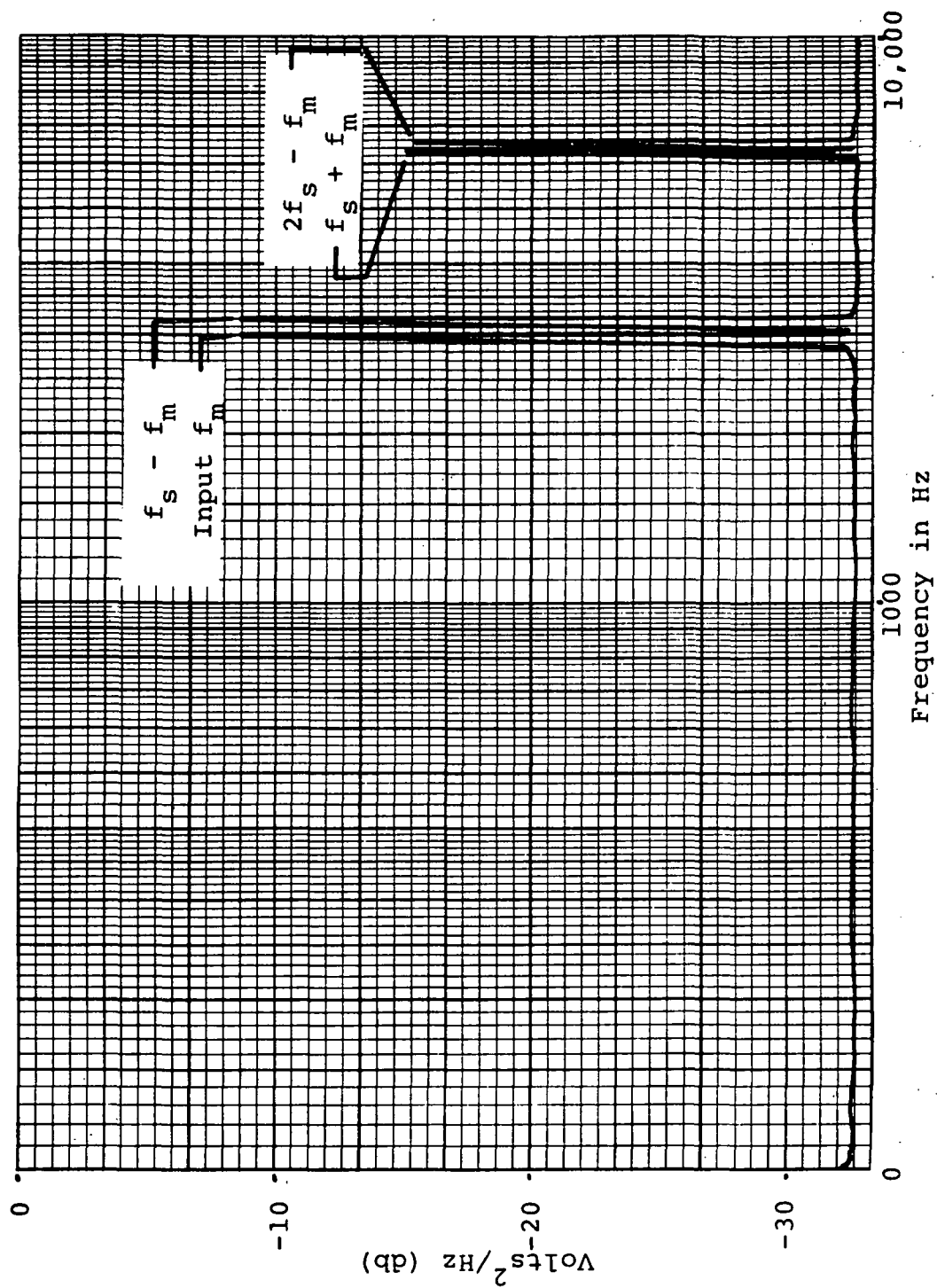


Fig. 4-5. 2000 Hz Sine wave Sampled at 4150 Hz and Quantized with 32 Levels

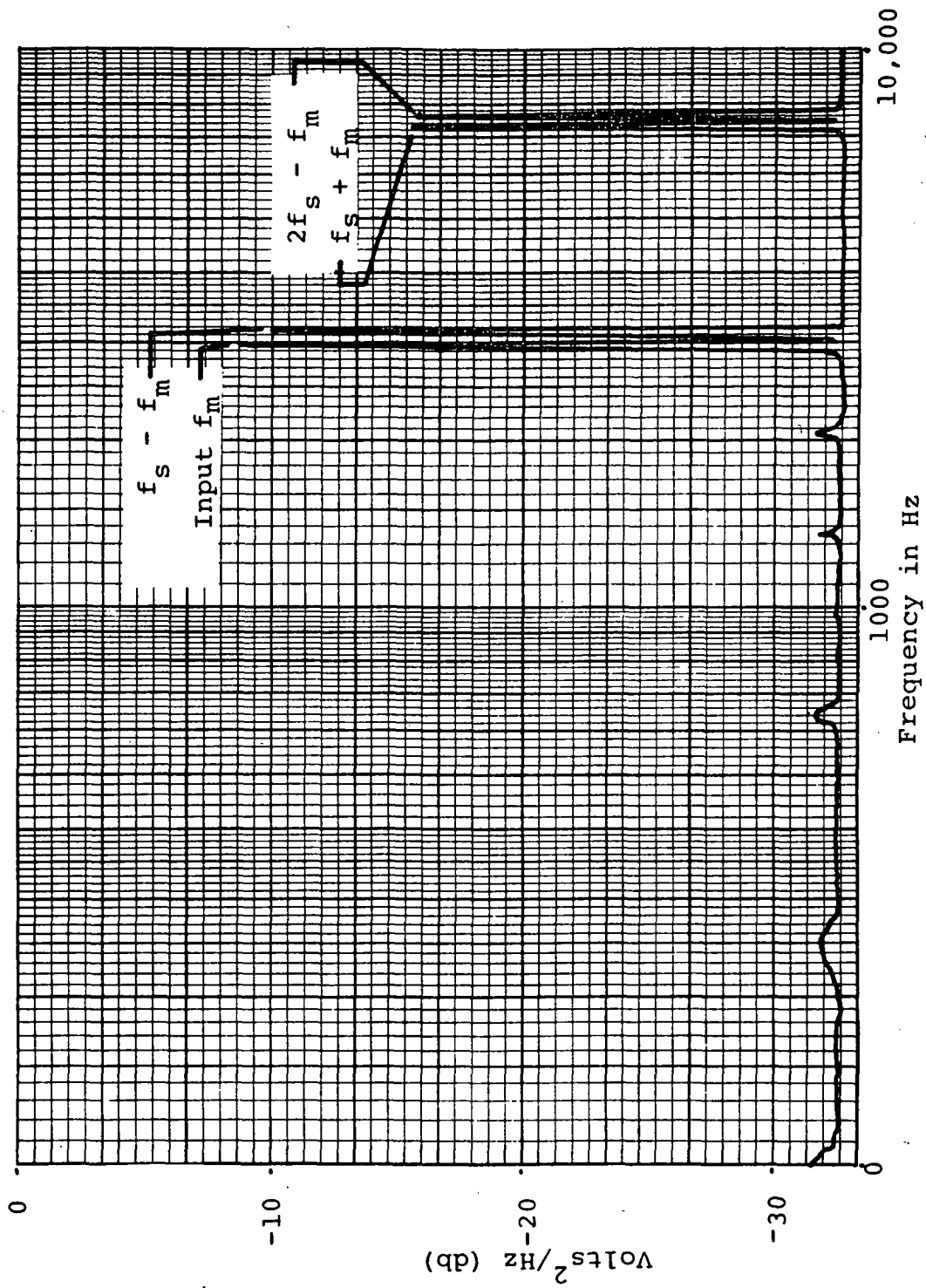


Fig. 4-6. 2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 16 Levels

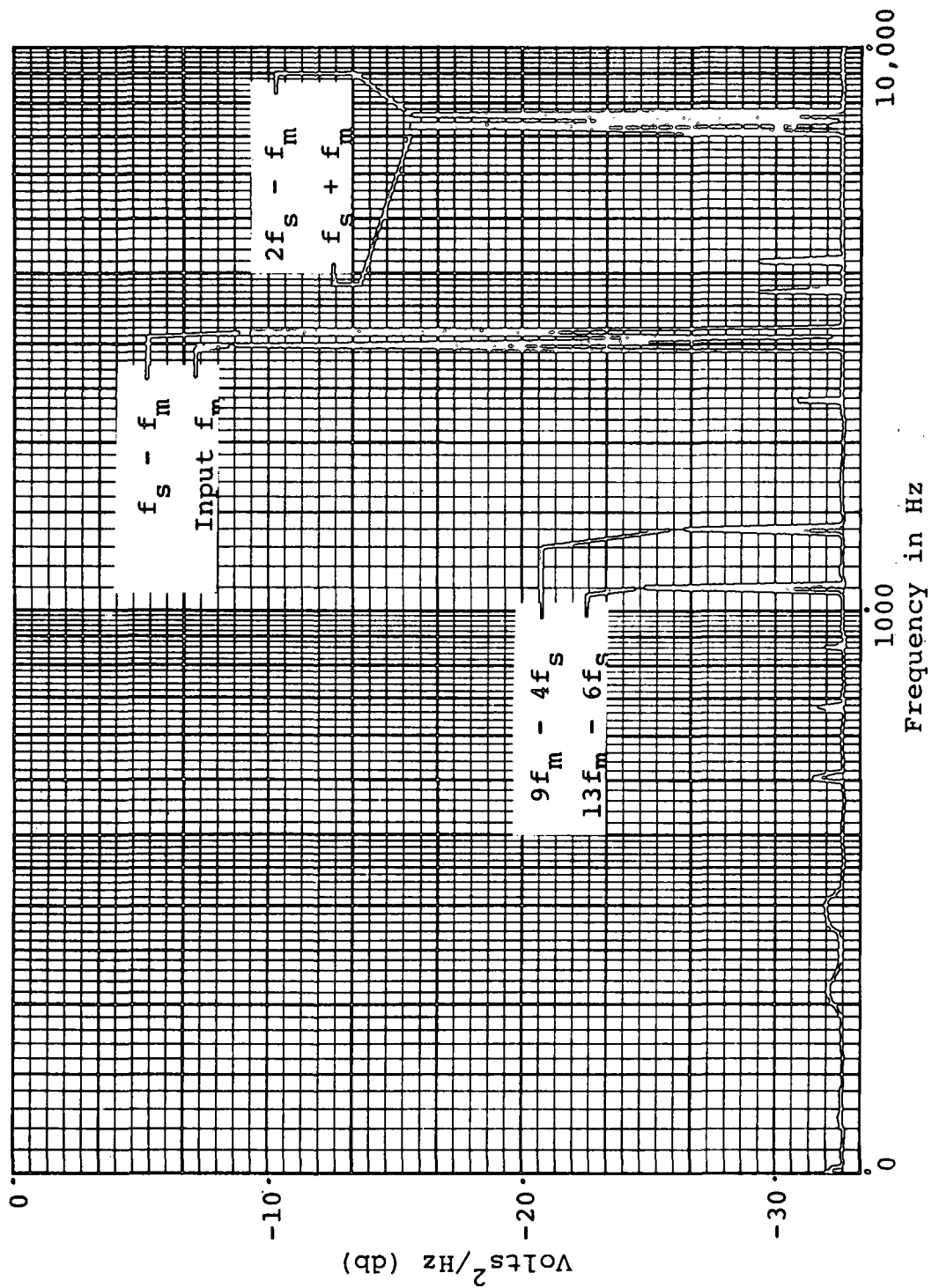


Fig. 4-7. 2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 8 Levels



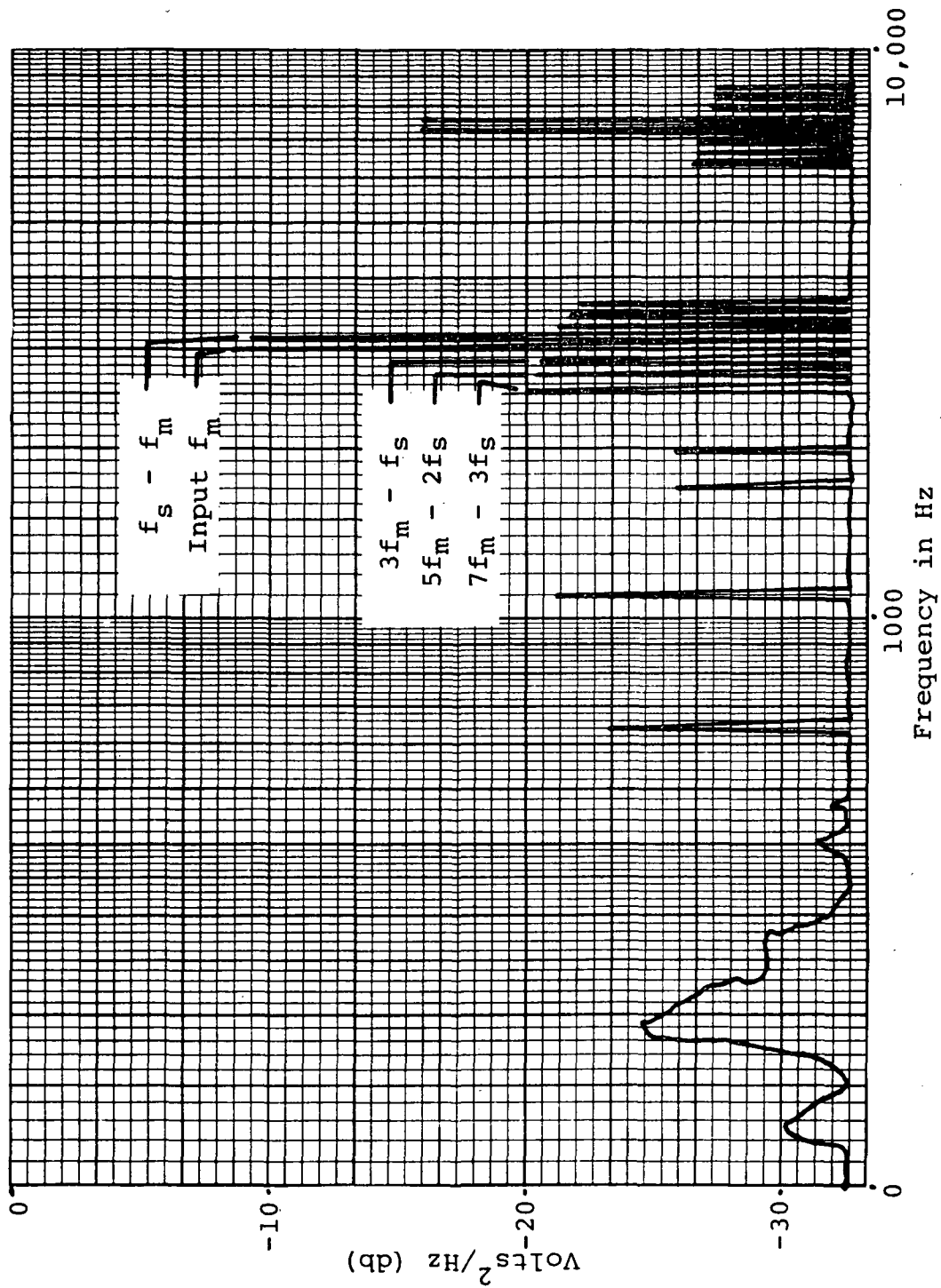
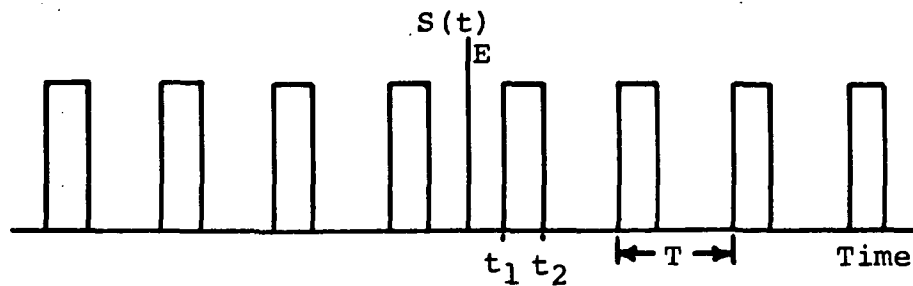
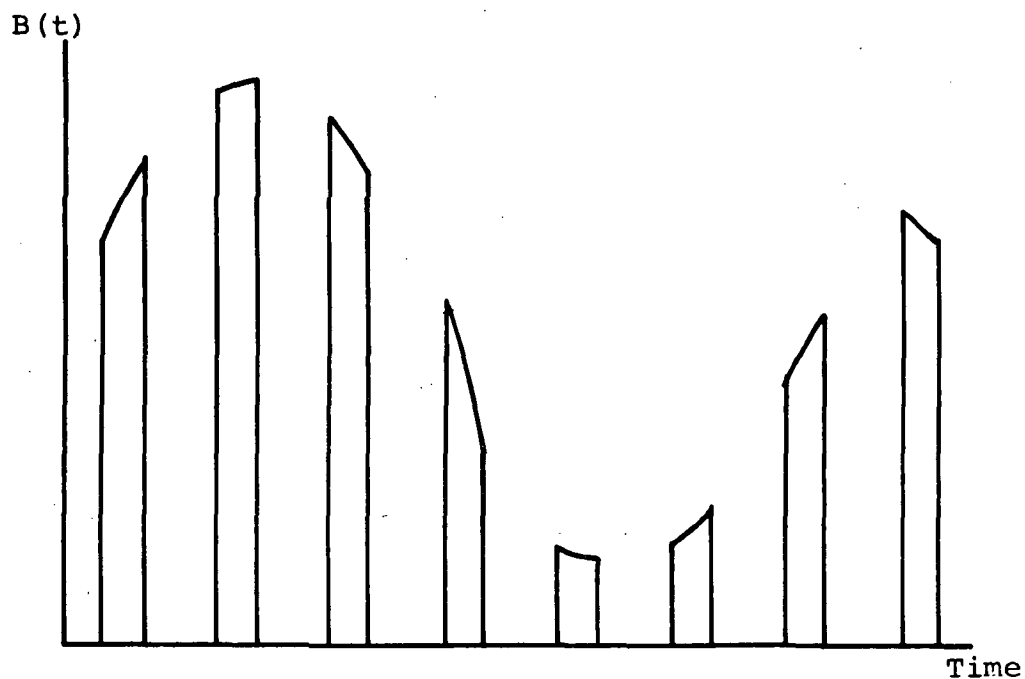


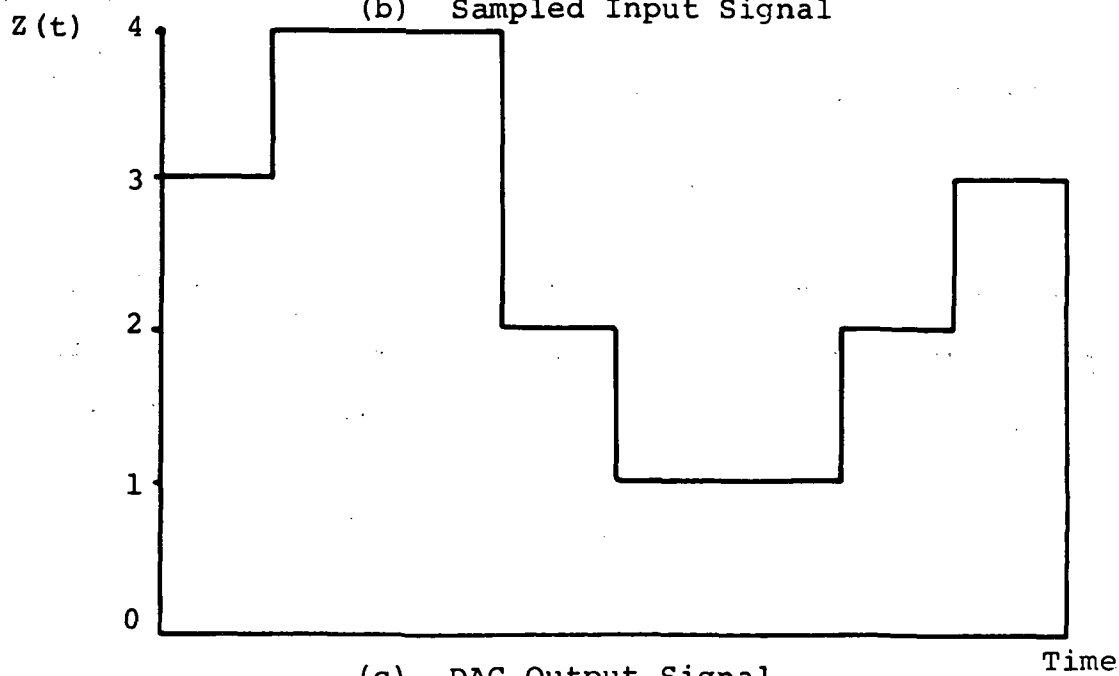
Fig. 4-8. 2000 Hz Sinewave Sampled at 4150 Hz and Quantized with 4 Levels



(a) Sampling Function



(b) Sampled Input Signal



(c) DAC Output Signal

Fig. 4-9. ADC Sampling Function and DAC Approximating Function

function are given by  $C_h$  .

$$C_h = \int_{t_1}^{t_2} E e^{-jh\omega_0 t} dt \quad (4-10)$$

where  $\omega_0 = \frac{2\pi}{T}$

$$= \frac{E}{h\omega_0} \left( \frac{e^{-jh\omega_0 t_2} - e^{-jh\omega_0 t_1}}{-j} \right) \quad (4-11)$$

For  $t_2 - t_1 = \Delta t$  or  $t_2 = t_1 + \Delta t$  , we have

$$C_h = \frac{E}{h\omega_0} e^{-jh\omega_0 t_1} \left( \frac{1 - e^{-jh\omega_0 \Delta t}}{j} \right) \quad (4-12)$$

$$C_h = E\Delta t \frac{\sin h\omega_0 \Delta t/2}{h\omega_0 \Delta t/2} e^{-j(h\omega_0 t_1 + h\omega_0 \Delta t/2)} \quad (4-13)$$

or

$$f(t) = \frac{E\Delta t}{T} \sum_{h=-\infty}^{\infty} \frac{\sin h\omega_0 \Delta t/2}{h\omega_0 \Delta t/2} e^{j(h\omega_0 t - \phi)} \quad (4-14)$$

where  $\phi = h\omega_0(t_1 + \Delta t/2)$

If a sinusoidal signal  $I(t) = A \cos \omega_m t$  is sampled by our sampling function in the ADC, the  $C_h$  term in the Fourier series of the output waveform  $B(t)$  (see Fig. 4-9(b)) is given by

$$C_h = \int_{t_1}^{t_2} EA \cos \omega_m t e^{-jh\omega_s t} dt$$

for  $\omega_o = \omega_s =$  sample rate in radians/second

$$\begin{aligned} &= \frac{EA}{2} \left[ \frac{e^{j(h\omega_s - \omega_m)\Delta t/2} - e^{-j(h\omega_s - \omega_m)\Delta t/2}}{j(h\omega_s - \omega_m)} \right. \\ &\quad \left. + \frac{e^{j(h\omega_s + \omega_m)\Delta t/2} - e^{-j(h\omega_s + \omega_m)\Delta t/2}}{j(h\omega_s + \omega_m)} \right] \\ &= EA \frac{\Delta t}{2} \left[ \frac{\sin(h\omega_s - \omega_m)\Delta t/2}{(h\omega_s - \omega_m)\Delta t/2} + \frac{\sin(h\omega_s + \omega_m)\Delta t/2}{(h\omega_s + \omega_m)\Delta t/2} \right] \quad (4-15) \end{aligned}$$

$$B(t) = I(t) S(t)$$

$$= \frac{EA}{T} \frac{\Delta t}{2} \sum_{h=-\infty}^{\infty} \left[ \frac{\sin(h\omega_s - \omega_m)\Delta t/2}{(h\omega_s - \omega_m)\Delta t/2} + \frac{\sin(h\omega_s + \omega_m)\Delta t/2}{(h\omega_s + \omega_m)\Delta t/2} \right] e^{jh\omega_s t} \quad (4-16)$$

The digital-to-analog conversion process in the DAC approximates the ADC input signal  $I(t)$  by generating a staircase approximation function  $Z(t)$  with step amplitudes corresponding to the quantized sample amplitudes of the ADC waveform  $B(t)$  (see Fig. 4-9(c)). Therefore  $Z(t)$  will have frequency components similar to those found in  $B(t)$  since the two waveforms differ only in duty cycle.  $Z(t)$  can be

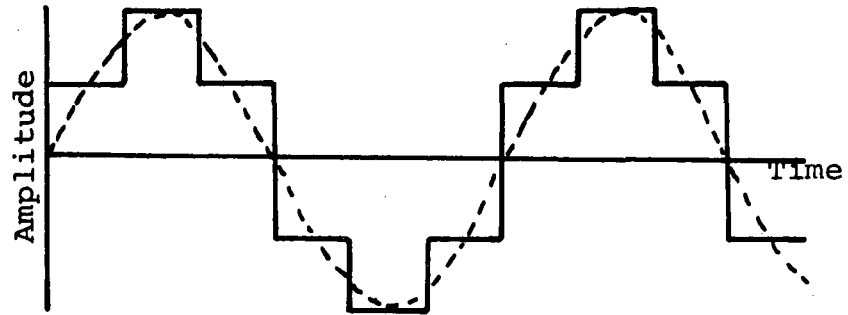
represented by the Fourier series for  $B(t)$  where the pulse width is increased to the limit of  $\Delta t = T$ . Therefore  $Z(t)$  is of the form

$$Z(t) = C \sum_{h=-\infty}^{\infty} \left\{ \frac{\sin[2\pi(hf_s - f_m)t]}{2\pi(hf_s - f_m)t} + \frac{\sin[2\pi(hf_s + f_m)t]}{2\pi(hf_s + f_m)t} \right\} e^{j2\pi hf_s t} \quad (4-17)$$

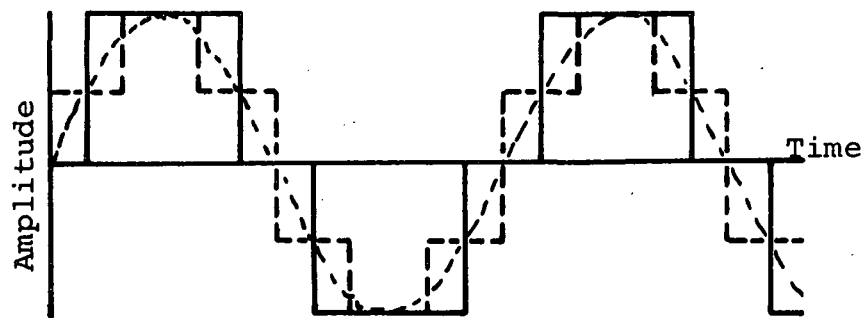
where  $f_s$  = the sampling frequency, and

$f_m$  = the frequency of the input signal.

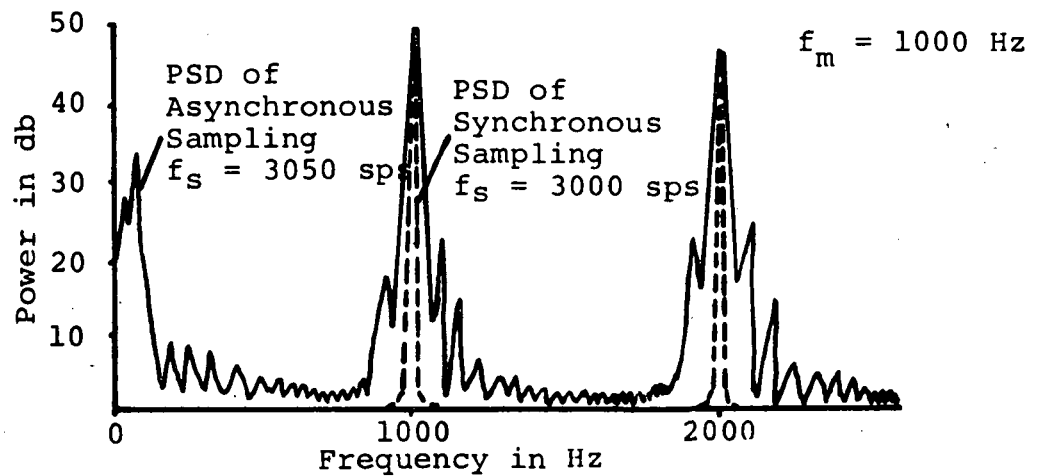
The amplitude of the harmonic components of the sum and difference components of  $f_s$  and  $f_m$  depend on the number of quantization levels used in the analog-to-digital conversion process. These components are also affected by the phase relationships between the sampling function and the input signal. Figure 4-10(a) shows the 4-level staircase DAC approximation function for a 1000 Hz sinewave sampled by a 3000 pps (pulse per second) sampling function that was synchronized with the input signal. The input sinewave was sampled in the same place for every consecutive cycle. The PSD plot (Fig. 4-10(c)) shows only the difference frequency  $f_s - f_m = 2000$  KHz and  $f_m = 1000$  KHz. When this same input signal,  $f_m = 1000$  Hz, is sampled asynchronously by a 3050 pps sampling function, consecutive DAC approximation waveforms are not the same for every cycle. When the sampled signal and the sampling function are asynchronous, much



(a) Synchronous Sampling



(b) Asynchronous Sampling



(c) Superimposed PSD's

Fig. 4-10. Sampling Function Synchronous and Asynchronous with ADC Input

quantization noise is generated (see Fig. 4-10(c)). The locus formed by the corners of the DAC approximation function when it is not synchronized to the input sinewave (normal case) produces the waveform shown in Fig. 4-11 when viewed on an oscilloscope. When the rate of sampling increases,  $\Delta t$  decreases and the error area decreases. When the number of quantization levels increases, the amount of uncertainty (or error) in the amplitudes of the steps of the DAC approximation function decreases. The unwanted harmonics generated by the quantization errors do not have a flat amplitude versus frequency distribution. As can be seen in Figs. 4-7, 4-8 and 4-10, the harmonic amplitudes increase as their frequencies approach that of the input signal or the sum and difference frequencies  $hf_s \pm f_m$ .

Much of the analysis of the sinewave outputs from a DAC can now be applied toward the analysis of speech spectrums. Speech is composed of many sinusoidal components. The vowel sounds are very sinusoidal in nature whereas the consonant sounds are more similar to white noise. Figure 4-12 shows the PSD of the phrase, "Top dog it's necessary to show you have heard wasp," recorded at the output of the DAC. The input speech was sampled at 4000 samples per second and 64 quantization levels were used to digitize it. Note the high energy content of the vowel formants centered around 500 Hz. This speech was very intelligible. Tapes prepared for

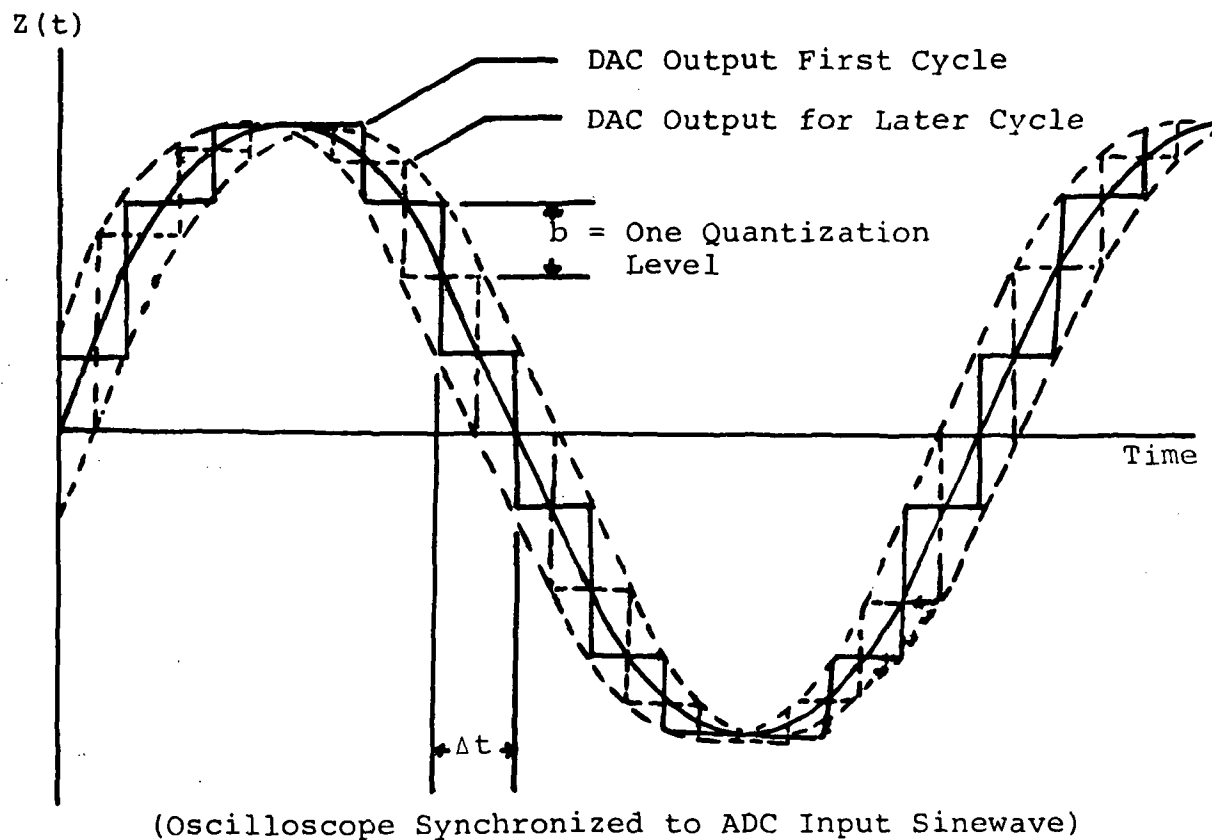


Fig. 4-11. Oscilloscope Waveform at Sampling Function Asynchronous with ADC Input (DAC Output)



Phrase, "Top Dog, it's necessary to show you have heard wasp," quantized with 64 levels

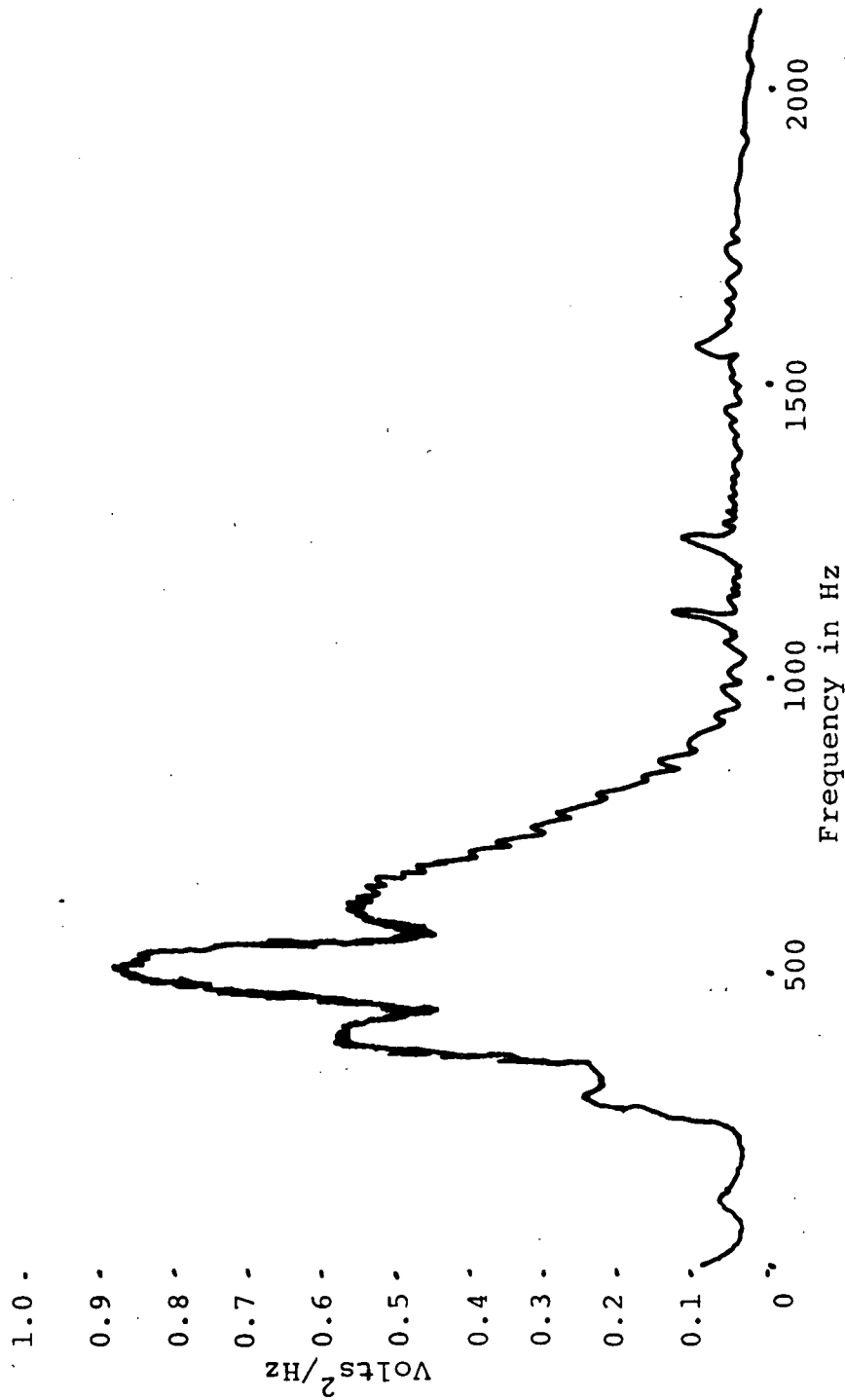


Fig. 4-12. PSD of DAC Output for Speech Sampled at 4000 Samples per Second

quantitative evaluation provided an average WI (word intelligibility) of 89 percent (see Appendix A for additional information on word intelligibility scoring and the significance of percent WI scores). If speech is sampled at rates less than 4000 samples per second, the first difference components ( $f_s - f_m$ ) of the vowel and consonant sounds are folded back on the input consonant and vowel sounds. These "harmonic distortion" products mask the desired speech components and seriously degrade the speech intelligibility. As the sample rate is decreased below 4000 samples per second, the consonant sounds are affected first. At a sample rate of 3000 samples per second, the consonant sounds from 1500 to 2000 Hz are being folded back on themselves. At a sample rate of 2500 samples per second, (Fig. 4-13), the high energy vowel sounds are being superimposed on the low energy consonant sounds between 1500 and 2000 Hz, causing a serious degradation of the speech intelligibility (WI = 69%). Therefore, the speech should be sampled at a minimum rate of 4000 samples per second and filtered such that no speech components with frequencies greater than 2000 Hz are sampled. Ideal lowpass filters are not realizable so a compromise selection of the maximum upper frequency components (hence, the minimum sample rate) must be made.

Assuming that the power spectrum for the consonant

Phrase, "Top Dog, it's necessary to show you have heard wasp,"  
quantized with 64 levels

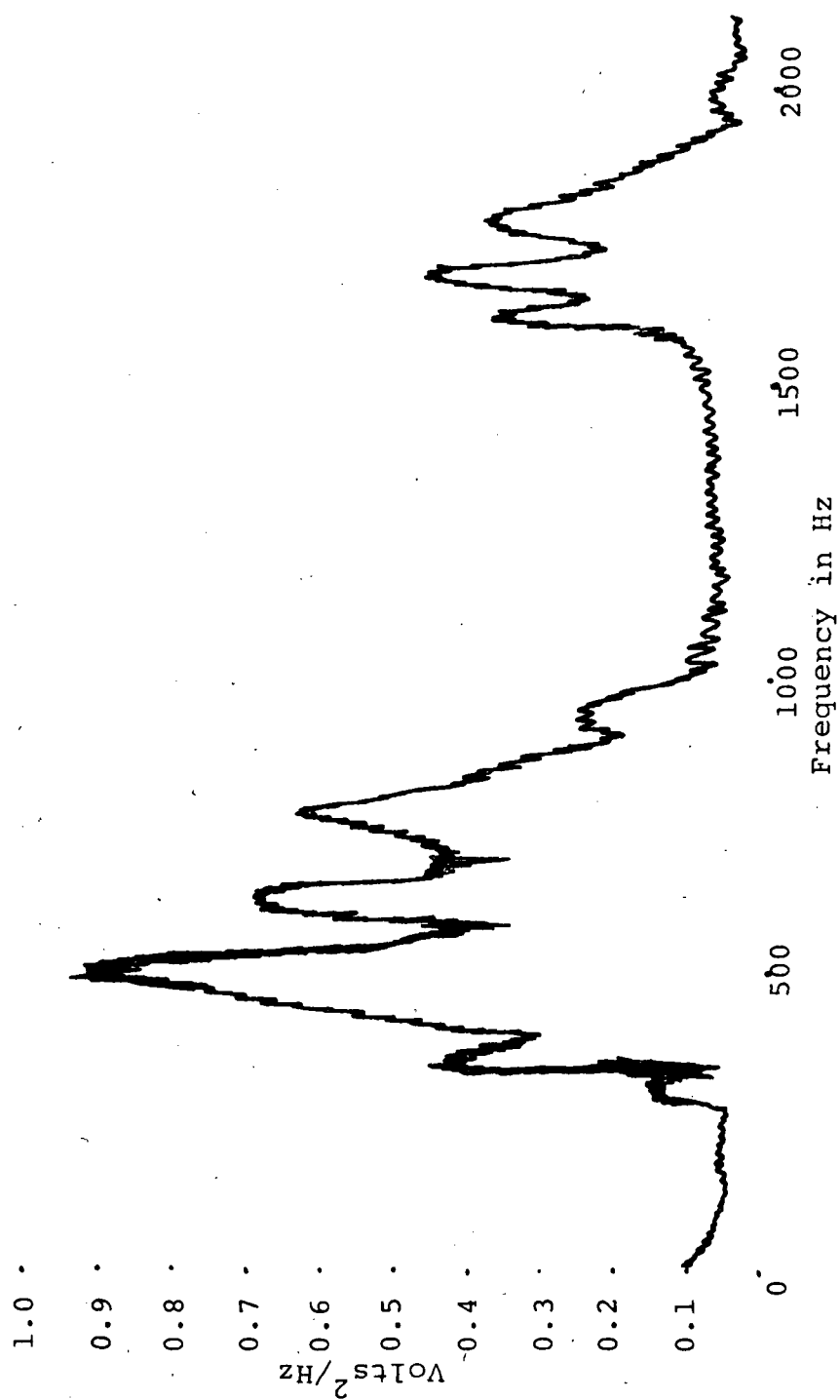


Fig. 4-13. PSD of DAC Output for Speech Sampled at 2500 Samples per second

sounds are flat, the spectrum of the "knee" of the lowpass filter output is given by

$$G^2(f) = \frac{A}{1 + (f/f_o)^{2m}} \quad (4-18)$$

where  $A$  = the power density of the flat portion of the spectrum,

$f_o$  = the half power frequency, and

$m$  = the rate of spectrum cutoff, i.e.,  $m = 1$  corresponds to 6 db per octave and  $m = 2$  corresponds to 12 db per octave (see Fig. 4-14(a)).

All of the difference components  $hf_s - f_m$  are bounded by the frequency  $f_s/2$  and the curve for  $m = 1$ . For frequencies greater than  $f_o$ , the spectrum is approximately

$$G^2(f) \approx A \left( \frac{f_o}{f} \right)^{2m} \quad (4-19)$$

and the energy in the difference components is

$$\begin{aligned} V_a &= \int_{f_s/2}^{\infty} G^2(f) df = \int_{f_s/2}^{\infty} A (f_o/f)^{2m} df \\ &= \frac{2^{2m-1} A f_o}{2m - 1} \left( \frac{f_o}{f_s} \right)^{2m-1} \end{aligned} \quad (4-20)$$

To compare the energy of the difference components with

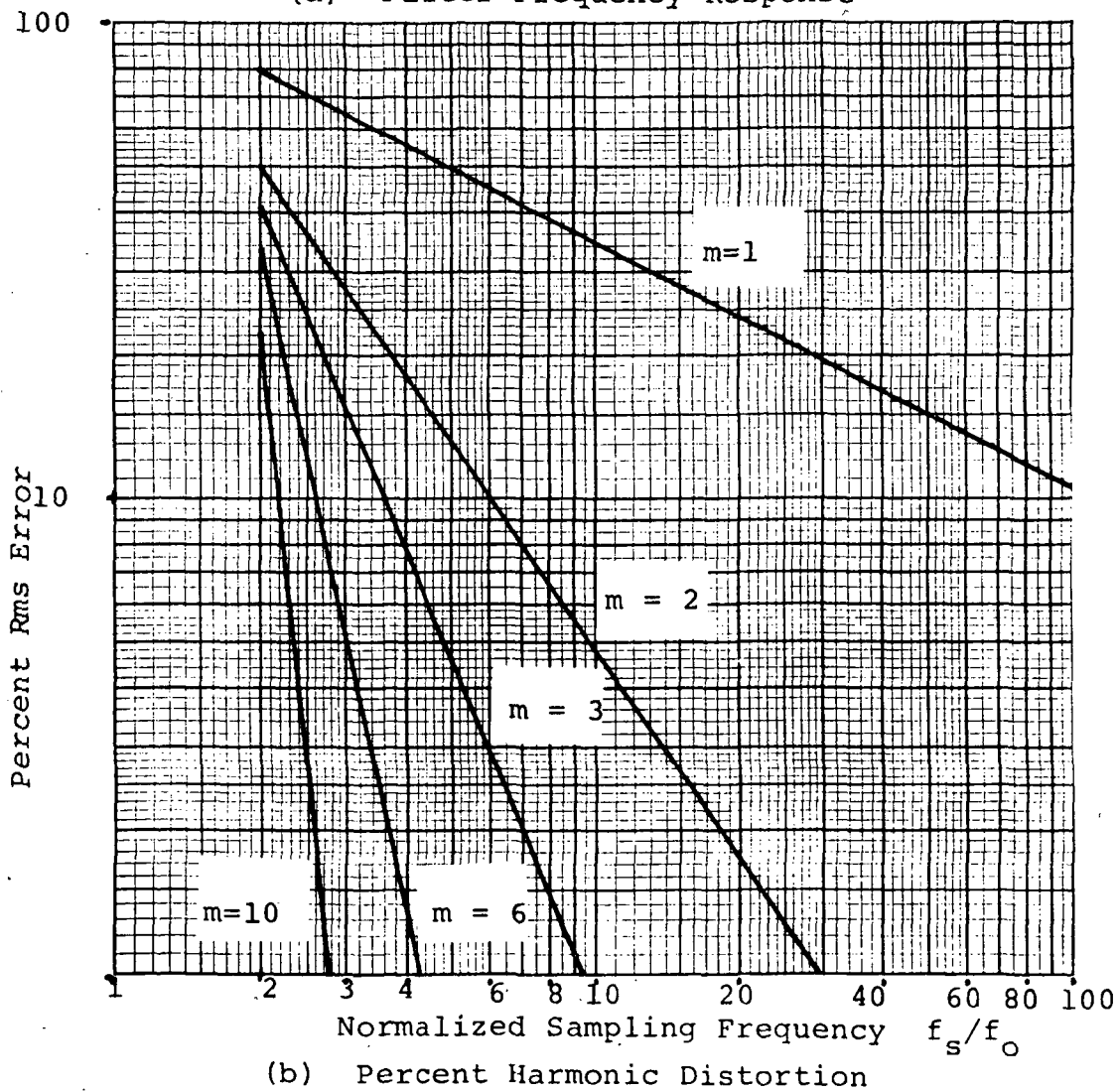
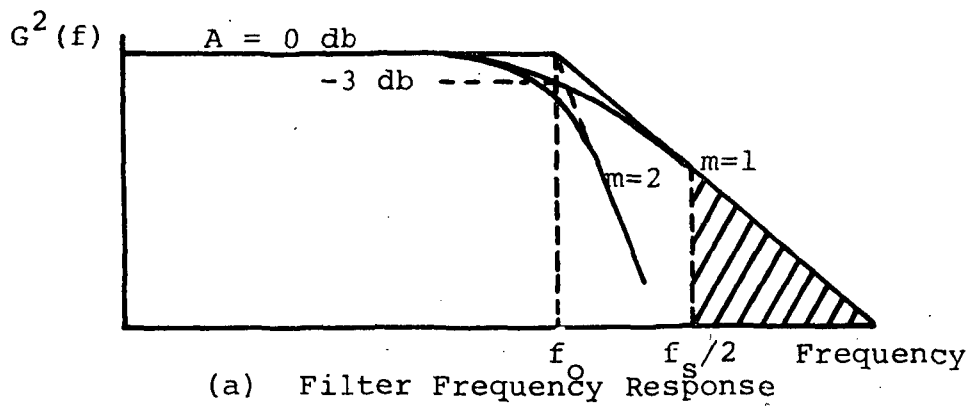


Fig. 4-14. Optimizing the Lowpass Filter Cutoff Frequency

that of the total speech power, we must first determine the amount of power in the total speech spectrum. A conservative estimate of the total speech power may be obtained by assuming that the speech spectrum is flat with the maximum energy versus frequency distribution equal to that of the consonant sounds. The total speech spectrum power is given by

$$\begin{aligned} V_s^2 &= \int_0^{\infty} G^2(f) df = \int_0^{\infty} \frac{A}{1 + (f/f_o)^{2m}} 2m df \\ &= \frac{\pi A f_o}{2m} \operatorname{cosec} \left( \frac{\pi}{2m} \right) \end{aligned} \quad (4-21)$$

The percentage of harmonic distortion caused by the frequency components greater than  $f_s/2$  is given by

$$V_p = \sqrt{\frac{V_o^2}{V_s^2}} = 2^m \left[ \frac{m}{\pi (2m-1)} \left( \frac{f_o}{f_s} \right)^{2m-1} \sin \frac{\pi}{2m} \right]^{1/2} \quad (4-22)$$

This quantity is plotted for five values of  $m$  in Fig. 4-14(b). Lowpass filters with an amplitude versus frequency rolloff of 36 db per octave over the cutoff frequency represent a practical limit for audio circuits. Filters with steeper rolloff characteristics become bulky and expensive. With a 36 db per octave filter and a half-power frequency  $f_o$  of 2000 Hz, the minimum ratio of sample rate  $f_s$  to  $f_o$  is approximately 2.5

to keep the harmonic distortion below 10 percent. The maximum value of 10 percent distortion was chosen since this is the level of distortion where the unwanted harmonics are not only noticeable but are beginning to degrade the channel's quality (Schmidt, 1968). Therefore, the minimum sampling frequency is 5000 samples per second for a bandpass filter with a 3 db cutoff frequency of 2000 Hz and a rolloff of 36 db per octave.

The minimum number of quantization levels necessary to achieve highly intelligible voice communications over a digital channel is shown in Fig. 4-2. The maximum theoretical output S/N ratios obtainable for several quantization levels are given in this figure. It is assumed that the digital channel must perform with 100 percent sentence intelligibility. This is equivalent to approximately 90 percent word intelligibility. From previous test results of voice communication channels, it is known that a channel's output speech-to-noise ratio should be in excess of 15 db to obtain a word intelligibility score of 90 percent or above (Hirsh, 1969). If one chooses 8 quantization levels, the maximum output S/N ratio is approximately 15 db. However, 8 levels gives no margin in output S/N and a WI of 90 percent cannot be assured. A better choice is 16 levels. Sixteen quantization levels provide a maximum output S/N of approximately 20 db assuming the input S/N is in excess of 20 db. In summary, our minimum

calculated parameters for a digital voice channel with 100 percent sentence intelligibility (assuming no uncorrectable channel degradation) are: (1) ADC input and DAC output bandpass filters with a cutoff frequency of 2000 Hz and a rolloff of 36 db per octave, (2) a sample rate of 5000 samples per second, and (3) 16 quantization levels.

Laboratory tests were performed to verify these theoretical parameters (Culver, 1969). The ADC and DAC were operated at several sample rates and quantization levels. Figure 4-15 shows the results of these tests. Note that the previously selected parameters would have resulted in WI scores very close to the desired 90 percent. More intelligible output speech can be provided if the sample rate is increased to 6000 samples per second. Increasing the number of quantization levels from 16 to 32 at a sample rate of 5000 samples per second does not give any appreciable improvement. The new lowpass filter cutoff frequency is now  $6000/2.5 = 2400$  Hz. Therefore, both experimental and calculated results support the conclusion that the minimum parameters for 100 percent sentence intelligibility (assuming no uncorrectable channel degradation) are: (1) ADC input and DAC output bandpass filters with a cutoff frequency of 2400 Hz and a rolloff of 36 db per octave, (2) a sample rate of 6000 samples per second, and (3) 16 quantization levels.



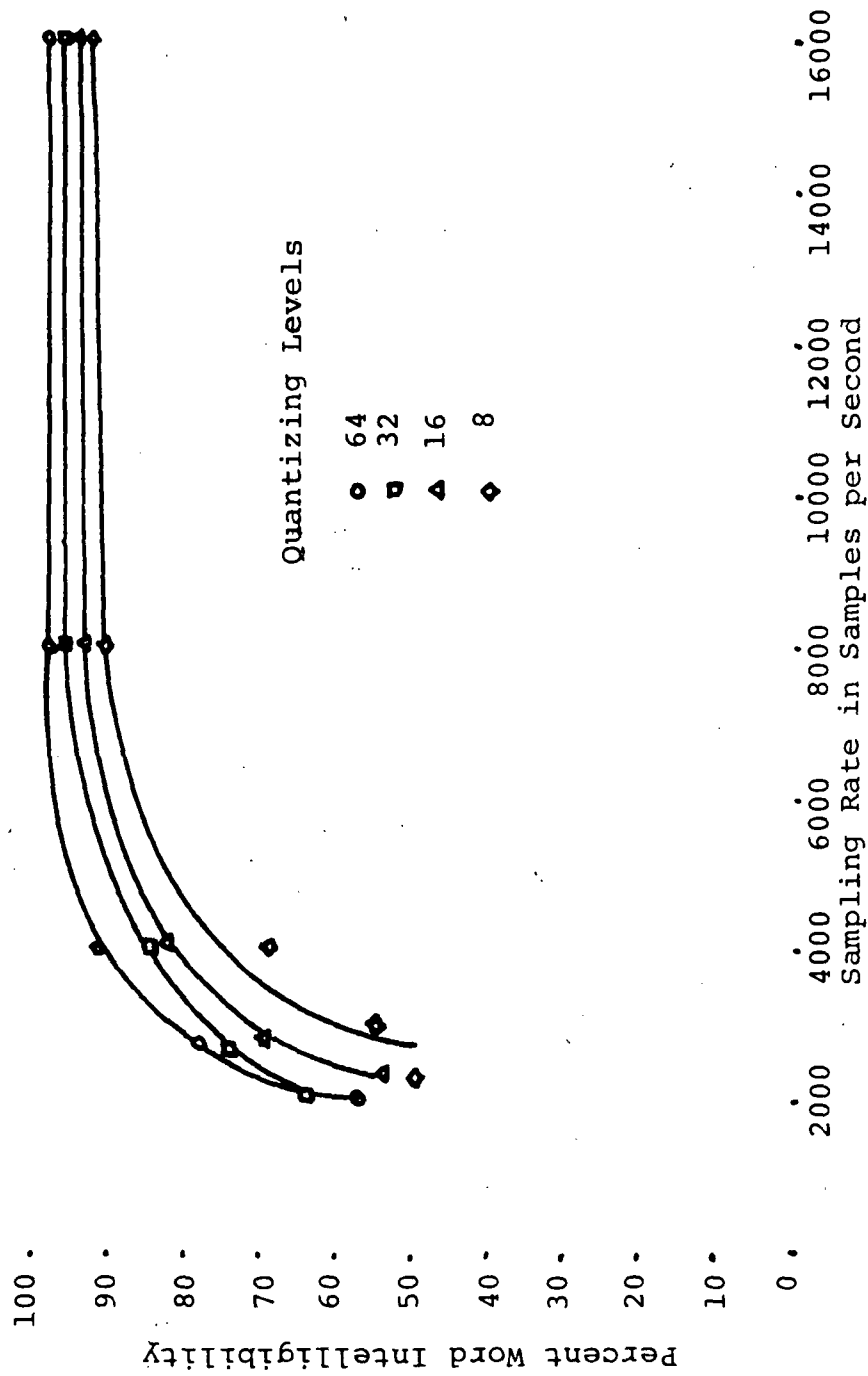


Fig. 4-15. Sample Rate Versus W.I. for Speech Quantized with 8, 16, 32 and 64 Levels

Power spectral distribution plots of speech with quantization noise show that consonant masking becomes very significant for 8 or less quantization levels. Figure 4-16 shows the PSD for the input speech phrase, "Top dog, it's necessary to show you have heard wasp." Note the vowel sound formant peaks from 0 to 1000 Hz and the consonant energy between 1000 and 3000 Hz. When this phrase is quantized using 32 and 16 levels, the DAC output does not contain any noticeable quantization noise (see Figs. 4-17 and 4-18). By comparing the two plots for 16 and 8 quantization levels (Figs. 4-18 and 4-19), it can be seen that the noise has significantly increased the total power in the frequency range containing the consonant frequencies for 8 quantization levels. When 4 levels are used (Fig. 4-20) the energy in the frequency band of 1000 to 3000 Hz is nearly twice the amount measured when 16 levels were used. This much noise superimposed on the consonant sounds requires the listener to recognize words and syllables based on vowel sounds and context which usually results in a low probability of correct message perception.

Phrase, "Top dog, it is necessary to show you have heard wasp."

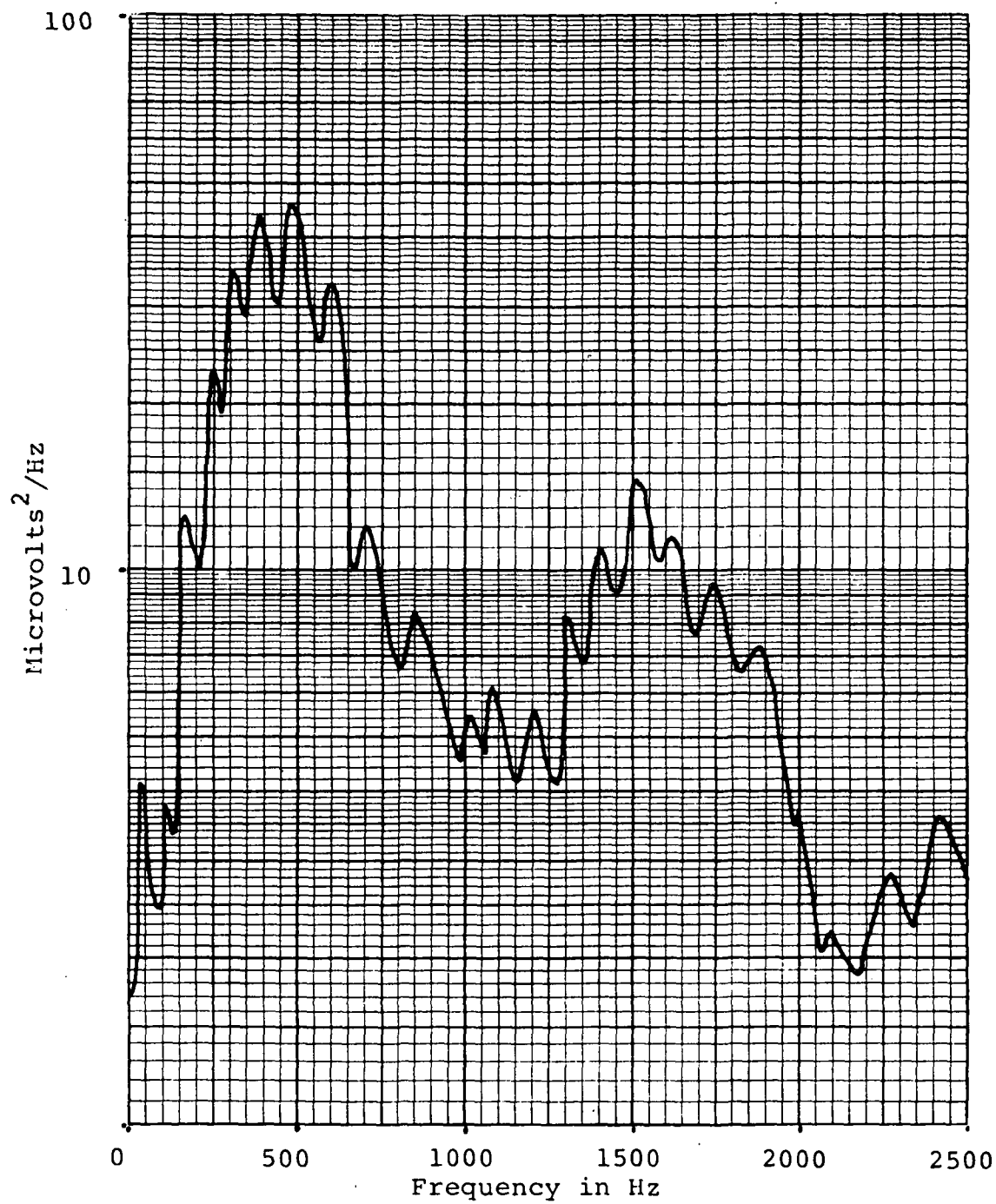


Fig. 4-16. PSD of ADC Input

Phrase, "Top dog, it is necessary to show you have heard wasp," sampled at 6700 samples per second

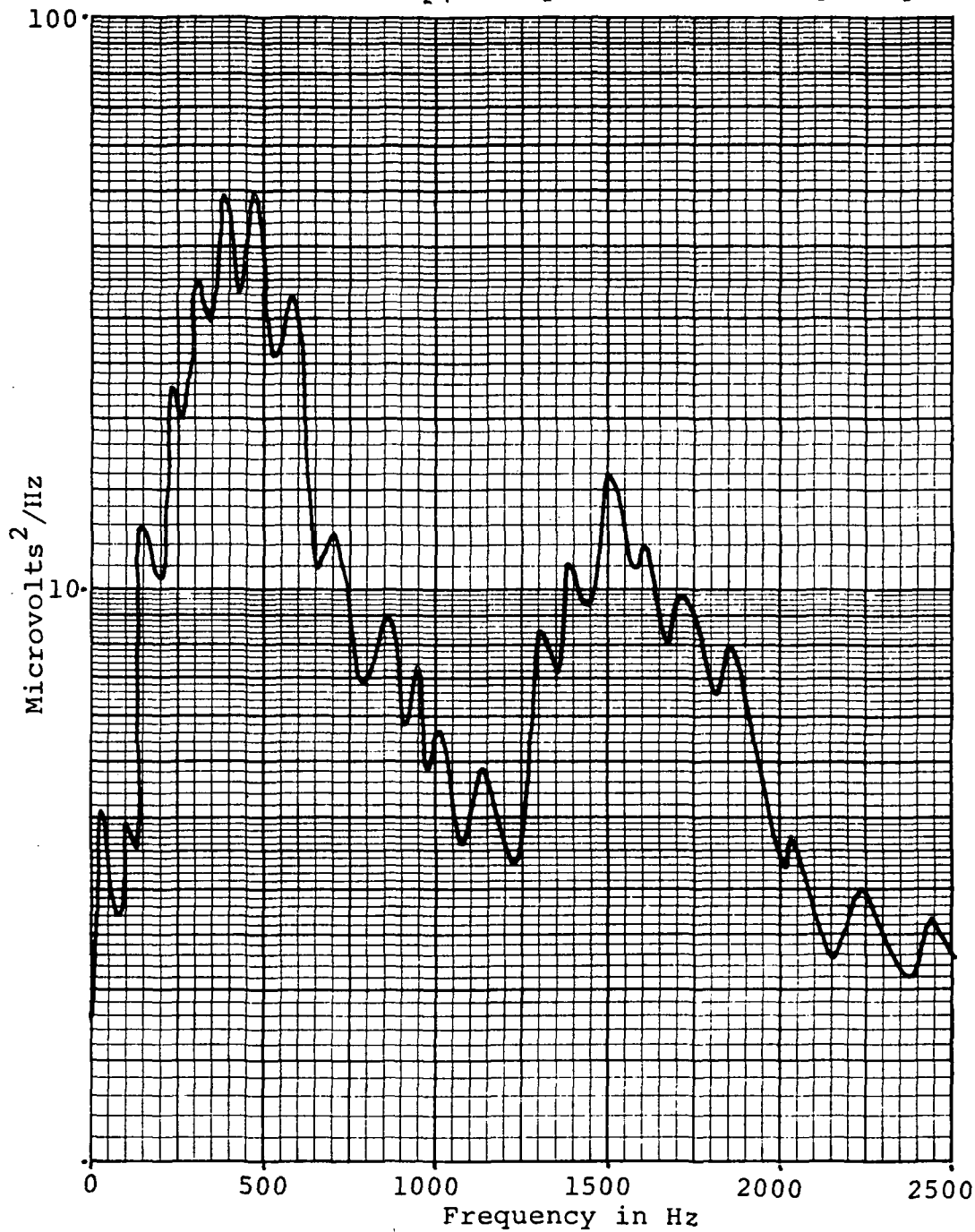


Fig. 4-17. PSD of DAC Output for Speech Quantized with 32 Levels

Phrase, "Top dog, it is necessary to show you have heard wasp," sampled at 6700 samples per second.

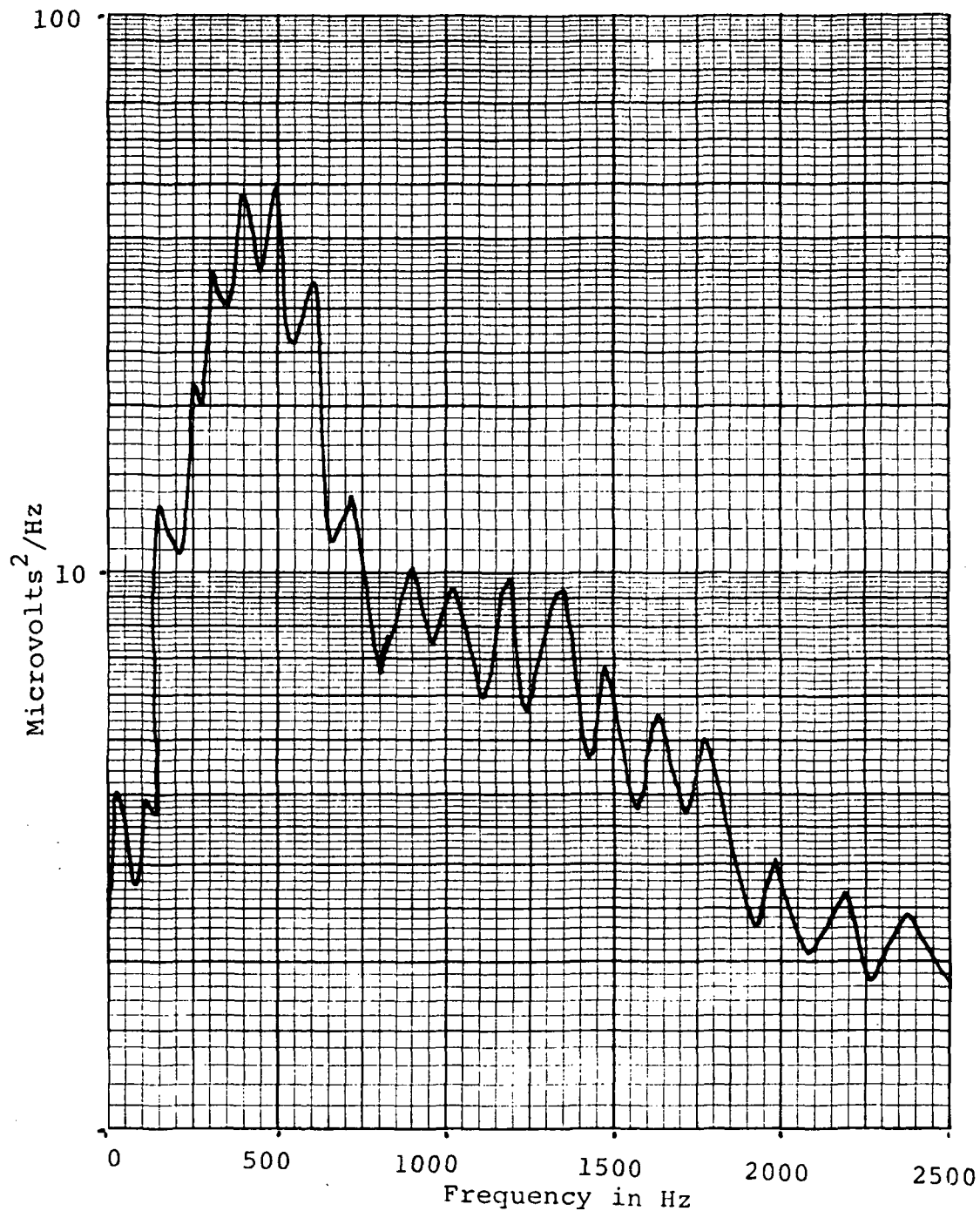


Fig. 4-18. PSD of DAC Output for Speech Quantized with 16 Levels

Phrase, "Top dog, it is necessary to show you have heard  
wasp," sampled at 6700 samples per second

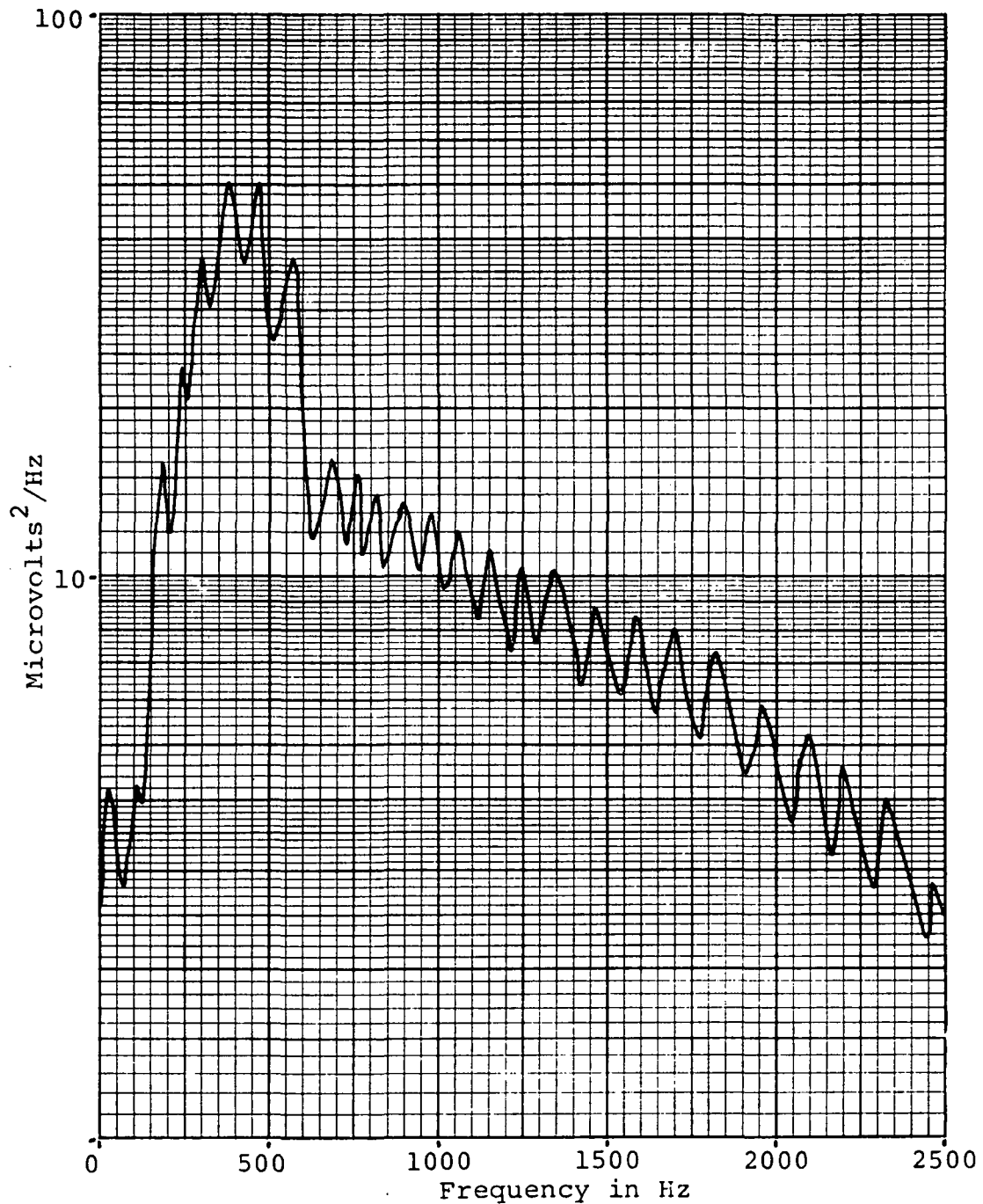


Fig. 4-19. PSD of DAC Output for Speech Quantized  
with 8 Levels

Phrase, "Top dog, it is necessary to show you have heard wasp," sampled at 6700 samples per second.

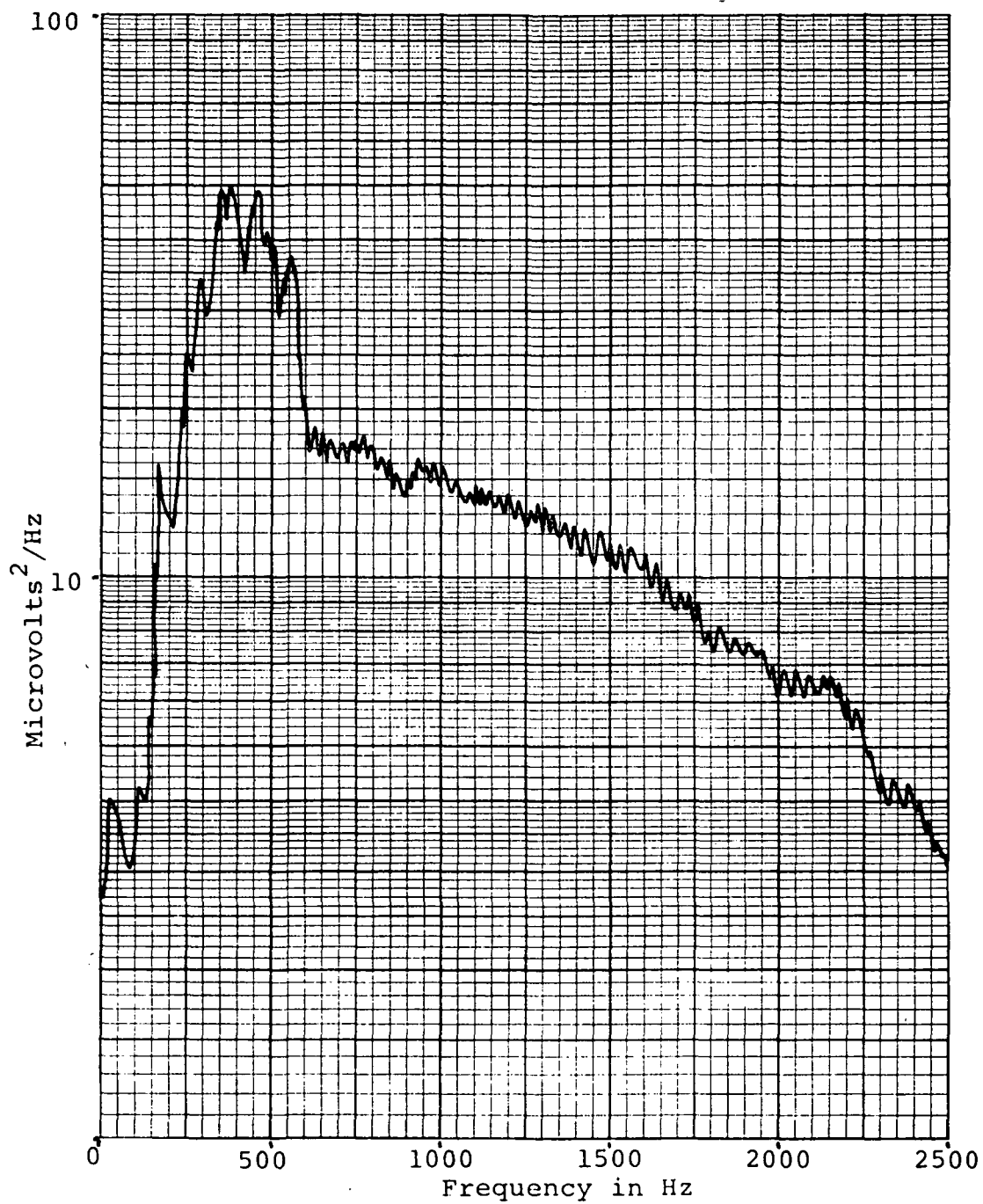


Fig. 4-20. PSD of DAC Output for Speech Quantized with 4 Levels

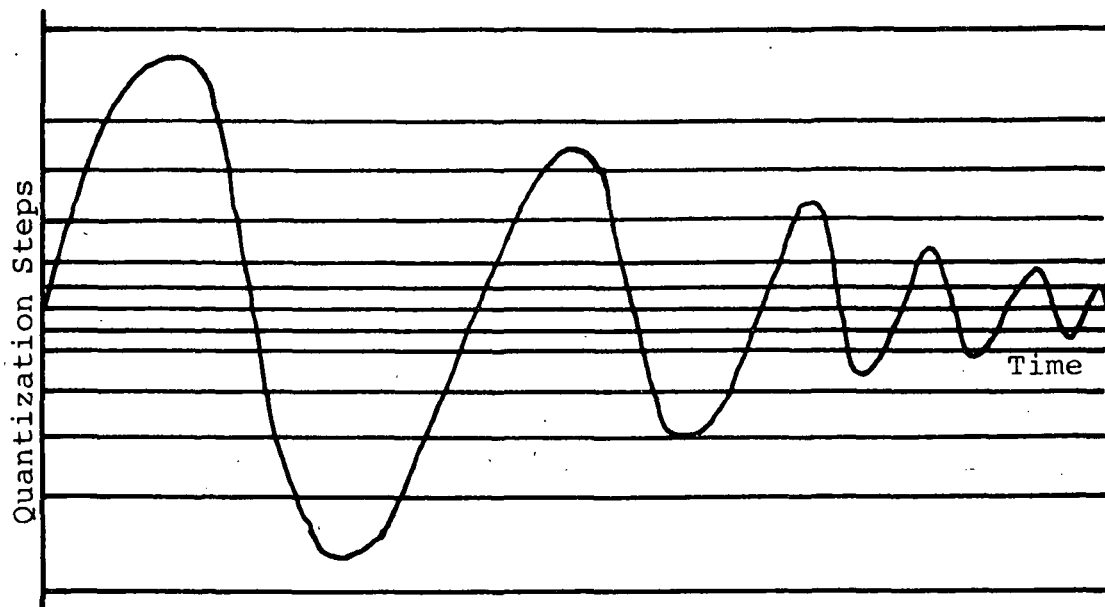
## CHAPTER V

### MINIMIZING QUANTIZATION NOISE

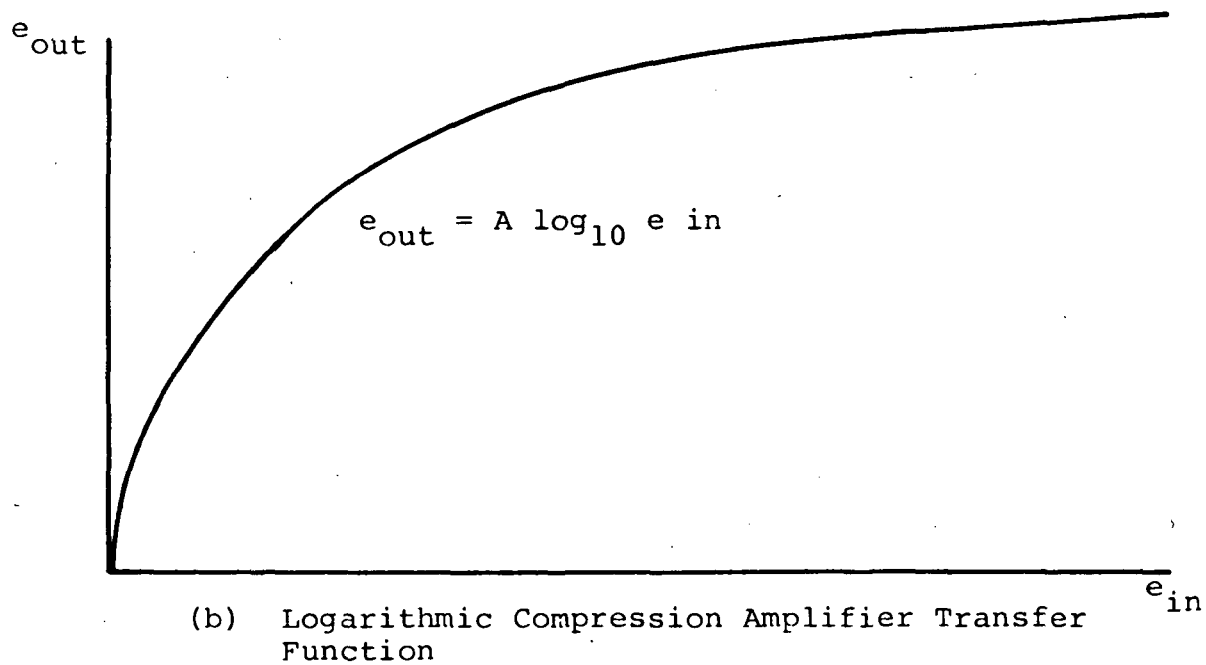
Various techniques to minimize quantization noise in a digital voice communication channel are discussed in this chapter. The first approach is trivial, namely, do not quantize the sampled speech with any less than 16 levels. However, if the overall channel bandwidth must be minimized, there are techniques that can be employed to reduce the overall bit rate by reducing the number of quantization levels required before quantization noise becomes excessive. A common approach is nonlinear quantization.

Nonlinear quantization employs unequal quantization steps. The steps are tapered to give fine divisions for low amplitude signals and course divisions for high amplitude signals (see Fig. 5-1(a)). A complimentary circuit is provided in the DAC to emphasize the high amplitude signals and de-emphasize the low amplitude signals to make the overall combination linear. For speech, the input signal is normally passed through a compression amplifier prior to digitizing. The output of the DAC is passed through an expansion network. A typical transfer function curve for the compression amplifier is shown in Fig. 5-1(b). Low amplitude consonant sounds are amplified while the high amplitude vowel sounds are attenuated in this type of





(a) Nonlinear Quantization



(b) Logarithmic Compression Amplifier Transfer Function

Fig. 5-1. Analog Nonlinear Encoding

amplifier. When the output of the compression amplifier is digitized, the consonant sounds receive a weighting more equal to that given the vowel sounds in a linear system. Consequently, when less than 16 quantization levels are used to digitize the speech, the relative ratio of quantization noise to the desired signal is more constant for both the consonant and vowel sounds. This reduces the effect of quantization noise at the expander output by increasing the ratio of consonant amplitudes to quantization noise amplitudes. Therefore, with nonlinear encoding the consonant and vowel sounds are more equally affected by quantization noise whereas the consonant sounds are affected first in a linear encoding system.

A laboratory experiment was conducted to determine the effects of analog companding (compression/expansion) the speech in conjunction with a linear ADC and DAC. The WI curves for 32, 16, 8 and 4 quantization levels are shown in Fig. 5-2. Note that the minimum parameters for achieving 90 percent WI with a linear system, namely, a sample rate of 6000 samples per second and 16 quantization levels, produce WI scores in excess of 95 percent WI when nonlinear encoding is employed. If nonlinear compression and expansion speech processing is used at the input of the ADC and the output of the DAC, the sample rate can be reduced to 5000 samples per second and the number of quantization levels

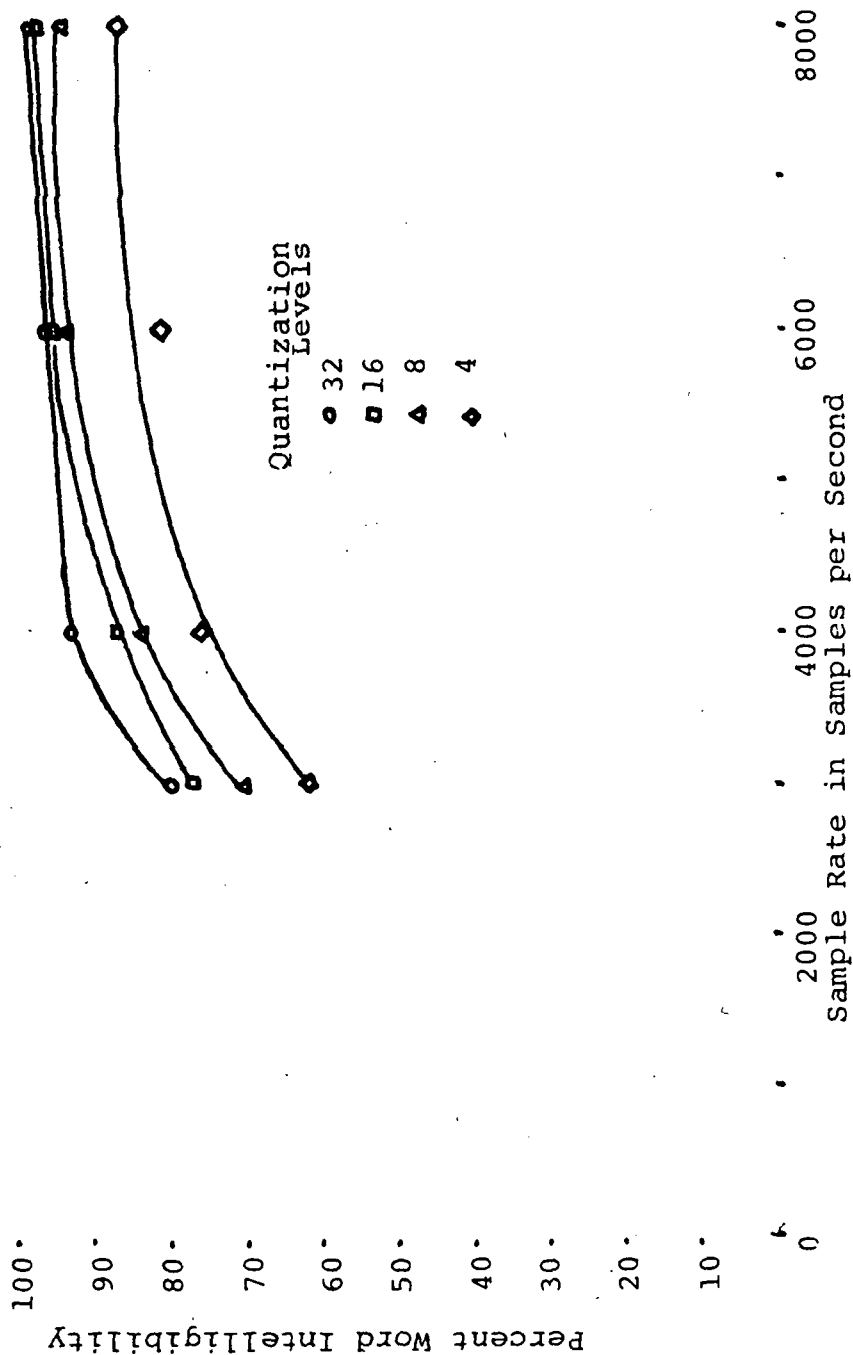


Fig. 5-2. Sample Rate Versus WI for Companded Speech Quantized at 4, 8, 16 and 32 Levels

can be reduced to 8 to achieve a 90 percent WI score for the channel. The lowpass filter cutoff frequency will also have to be reduced to  $5000/2.5 = 2000$  Hz.

Nonlinear quantization can also be achieved by employing tapered steps in the analog-to-digital conversion process. Each sample value of the input analog signal is compared to reference voltages within the ADC to determine the binary equivalent ( $\pm$  one-half of a quantization step). If these reference voltage levels were adjusted to give weighted binary equivalent values to the sample values, the high amplitude vowel sounds could be weighted downward and the low amplitude consonant sounds could be weighted upward. The results would be similar to those for the analog compression amplifier at the input of a linear ADC. A nonlinear DAC would have to be matched to the nonlinear ADC to obtain an overall linear system. Nonlinear ADC's and DAC's are not readily available and must therefore be built especially for voice transmission. Once an ADC or DAC has been modified for nonlinear speech encoding or decoding, it would probably not be suitable for use with any other data. It is for these reasons that speech, and other data, is nonlinearly processed in its analog form so that common linear ADC and DAC equipment can be used.

Another type of speech processing that can be employed to reduce the effects of quantization noise is pre-emphasis.

Recall that the consonant sounds are affected first by quantization noise and that they have the most power in the frequency band of 2000 to 3000 Hz. Therefore, if we increase the relative amplitude of the consonant sounds with frequencies in excess of 1000 Hz with respect to the vowel sounds below 1000 Hz, quantization noise would begin to degrade the speech at a lower ratio of rms speech-to-noise power. The simplest form of a pre-emphasis network is a capacitor in series with the ADC input. One capacitor will provide 6 db per octave of pre-emphasis. The most precise pre-emphasis network is an RC network in series with the input of the ADC to select the frequency at which the 6 db per octave of pre-emphasis begins. For example, if the RC network is designed for pre-emphasis beginning at 500 Hz, signals at 1000, 2000 and 4000 Hz would be emphasized 6, 12 and 18 db, respectively, with respect to the 500 Hz signal. A speech purist would apply the DAC output to a speech de-emphasis network with a negative 6 db per octave slope to compensate for the unnatural sound of the speech passed through a pre-emphasis network. However, when pre-emphasis is employed in most existing voice communication systems used by male speakers, de-emphasis is not employed. The resulting output speech sounds "high-pitched" but it is very intelligible and can be understood better in conditions of high ambient noise.

## CHAPTER VI

### CONCLUSIONS AND RECOMMENDATIONS

High sentence intelligibility can be achieved from a linearly encoded PCM voice channel with a minimum bit rate of 24,000 bits per second. The minimum parameters for 100 percent sentence intelligibility are: (1) analog-to-digital and digital-to-analog converter lowpass filters with a 3 db cutoff frequency of 2400 Hz and a rolloff of 36 db per octave, (2) a sample rate of 6000 samples per second, and (3) 16 quantization levels. If less than 16 levels are used to quantize the input signal, the quantization noise generated in the digital-to-analog converter will begin to seriously degrade the output speech. The first speech components to be affected by quantization noise are the consonant sounds. If the low amplitude consonant sounds are amplified with respect to the high amplitude vowel sounds, more quantization noise can be tolerated at the output of the digital-to-analog converter. This can be accomplished by passing the speech through a logarithmic compression amplifier before encoding it in the analog-to-digital converter. The consonant sound weighting is restored at the output of the digital-to-analog converter by an expansion amplifier. With nonlinear encoding, the number of quantization levels can be reduced to 8 and the sample rate can be reduced to

5000 samples per second. The lowpass filter cutoff frequency must also be reduced to 2000 Hz.

It is recommended that more work be done to develop a digital-to-analog converter that is adapted to speech. The analog signal generation circuits should be designed to generate sinusoidal functions by straight-line interpolation instead of by filtering a staircase function.

## BIBLIOGRAPHY

- Black, H. S. Modulation Theory. New York: D. Von Nostrand Co., Inc., 1953.
- Culver, L. "Digitized Speech Test (PCM) Data Package, NASD No. 642D-836547," Internal Report for NASA/MSC, Contract NAS 9-5191, November, 1969.
- Flanagan, J. L. Speech Analysis, Synthesis, and Perception. New York: Springer-Verlag, 1965.
- Fletcher, H. Speech and Hearing in Communication. New York: D. Van Nostrand Co., Inc., 1961.
- Hirsh, L. M. "Apollo Voice Intelligibility Report," Internal Report for NASA/MSC, Contract NAS 9-1261, 1969.
- Kleiner, R. T. "Automatic Speech Recognition," Bellcomm, Inc., Washington, D. C., February, 1969.
- Schmidt, O. L. "MSFN Signal Data Demodulator System Audio Filter Analysis," Internal NASA/MSC Report EB68-3711(U), December, 1968.
- Schwartz, M. Information Transmission Modulation and Noise. New York: McGraw Hill, 1959.
- Stitby, H. L. Aerospace Telemetry. New Jersey: Prentice Hall, 1961.
- Taki, Yasuo. "Fundamentals of PCM Communication Systems," Electronics and Communications in Japan, Vol. 49, 1966, pp. 8-15.



## APPENDIX A

### WORD INTELLIGIBILITY SCORING

The unit, "percent word intelligibility," is a quantity representing the percentage of words correctly recognized at the output of a voice channel from a list read into its input. There are many different WI (word intelligibility) measurement techniques, each with its own biases of the output scores. All techniques attempt to make a repeatable measurement independent of the human subject's emotions and previous training. Single-syllable (monosyllable) words are usually used to prevent the selection of the correct word from adjacent syllable context.

The WI measurement technique used to evaluate the different laboratory test conditions covered in this thesis was based on the Harvard PB (phonetically balanced) word technique. Fifty monosyllable PB words were selected for each word list to cover most of the phonemes of the English language. Each word was enunciated in a carrier phrase such as "Top dog, top dog, it is the word ant that you should record now," for the word "ant." The PB words and their carrier phrases were recorded and later played into the analog-to-digital converter of the simulated digital voice channel to determine the effects of changing specific parameters. The output of the digital-to-analog converter was

recorded and later scored by subjects. The percentage of correctly perceived words was the percent WI for the channel.

A WI score can be related to the percentage of sentences that could have been interpreted over the voice link under test. Since sentences contain a large amount of redundant information, the sentence intelligibility score will be greater than the word intelligibility score for each test condition. Word intelligibility scores account for the same factors that degrade voice channels used for conversational speech. Channel noise and distortion that would degrade a voice channel's quality will also decrease the WI score. However, WI measurements will not normally account for sounds such as echoes or tones which are annoying but do not degrade the channel's intelligibility. The WI measurement technique used for the tests in this thesis was repeatable and showed sufficient sensitivity to changes in the test parameter variables to allow optimization of the digital voice channel's performance.