AN ABSTRACT OF THE THESIS OF

Eric L. Zahl for the degree of Master of Science in Electrical and Computer Engineering presented on August 4, 1989.

Title: Resolution Enhancement of Analog-to-digital Converters through Computationally Simple Digital Filtering

Abstract approved: Redacted for privacy

Sayfe Kiaei

The resolution of analog-to-digital converters can be distinguished as absolute resolution, or average resolution. This study reviews average resolution enhancement techniques and proposes a method which is particularly applicable as a low-cost modification to a high-speed waveform acquisition system. This method uses oversampling combined with computationally simple digital filtering to enhance the average resolution of an analog-to-digital (A/D) converter, while maintaining an absolute resolution which is at least that of the unmodified system. The average resolution enhancement is approximately $\frac{1}{2}\log_2(N) - \frac{1}{4}$ bits, where N is the oversampling ratio. The simple digital filter is also shown to be useful in reducing alias errors when sample rates are reduced to near Nyquist rates. Additive dither signals are shown to be useful in maintaining expected resolution enhancement for certain input signals.

Resolution Enhancement of Analog-to-digital Converters through Computationally Simple Digital Filtering

by

Eric L. Zahl

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Completed August 4, 1989

Commencement June 1990

APPROVED:

Redacted for privacy

Professor of Electrical & Computer Engineering in charge of major

Redacted for privacy

Head of department of Electrical & Computer Engineering

Redacted for privacy

Dean of Graduate School

Date thesis is presented: August 4, 1989

ACKNOWLEDGEMENTS

The completion of this thesis was made possible by the ample assistance I have received.

This research was sponsored, in part, by Tektronix, Inc. of Beaverton, Oregon. I am very grateful to the company as a whole. I would especially like to thank Shiv Balakrishnan of Tektronix who suggested this research topic and encouraged Tek's support. Truly, the conception of this thesis is Shiv's. He has proved to be rich source of knowledge, both theoretical, and practical. Throughout this research, I have been inspired by Shiv's ability to be critical, while maintaining a personal sense of humility.

I would like to give sincere thanks to my adviser, Dr. Sayfe Kiaei, who oversaw my progress and found the time to become personally involved. His encouragement and direction enabled me to complete this thesis. Without his advice, I would continue to be struggling to find a focus within this topic.

I thank my parents, Jack and Ellen Zahl, for the years of nurturing they have given me. This thesis is one fruit of their effort in developing my sense of self-worth.

I thank my best friend and wife, Linda Zahl, for her companionship, and her unconditioned support and love.

I thank my wife's parents, Ezzat'u'llah and Bobbie Varjavandi, for their dedication to ensuring that their daughter and I both received a quality education.

My deepest sense of gratitude is due to the Creator of this vast universe. I thank God for the incredible Providence I have received, and, all to often, taken for granted.

TABLE OF CONTENTS

LIST OF FIGURES

Simulation results

Resolution Enhancement of Analog-to-digital Converters through Computationally Simple Digital Filtering

Chapter 1

Introduction

Analog-to-digital (A/D) converter resolution can be evaluated from two perspectives [5]: absolute resolution based on each sample considered independently (eg. peak error and integral nonlinearity), or average resolution based on an average of samples (eg. signal to noise ratio (SNR) and effective bits). Enhancing the absolute resolution usually involves improving the accuracy of the circuit technology and design. Enhancing the average resolution can be accomplished by signal processing techniques.

This thesis examines the use of high sampling rates combined with digital filtering to enhance the average resolution of analog-to-digital (A/D) conversion. There are several ways in which the resolution will be enhanced by such a system.

Under many circumstances the quantization error sequence is broadband. If the input signal is sampled at a rate which is greater than the Nyquist rate (oversampled), the error which is out of the signal band can be reduced by digital filtering. Another factor which encourages the use of oversampled ADCs is the ability to trade anti-aliasing analog filter complexity for digital filter complexity. When oversampling is not used, a front-end analog anti-aliasing filter is usually required. This lowpass filter must have a narrow transition band and offer high attenuation in the stop band. In an oversampled system, a digital filter can be used to suppress the out of band components. The sample rate can then be reduced to near-Nyquist rates by decimation. A simple analog anti-aliasing filter, often a single pole filter [18], can be used with the oversampled system for an efficient combination of analog and digital system complexity. Other advantages include the reduction of out-of-band system noise, and the improvement of integral linearity by reducing out-of-band harmonics.

1.1 Purpose and motivation

The purpose of this thesis is to propose and analyze an oversampled A/D conversion system to increase the resolution of digitized waveform samples output at near-Nyquist rates. The system should :

1) operate at output sample rates ranging from a few Kilosamples/second to hundreds of Megasamples/second,

2) be a low-cost modification to a high-speed quantizer,

3) maintain the absolute error of the raw quantizer for all cases of signals where the output rate is at least the Nyquist rate.

These requirements are derived from the design of a general waveform acquisition system.

To meet the objectives of this thesis, a family of computationally simple digital lowpass filters is proposed which requires no multiplication. These filters enhance A/D resolution by reducing the quantization error out of the signal band.

1.2 Review of literature

Oversampled A/D converters can be distinguished by the absence or presence of feedback around the quantizer. Those methods which do not use feedback include:

1) lowpass filtering,
2) the method of Claasen et al. [8],
3) the method of Belcher [4].

The most popular methods which use feedback can be classified as [24]:
4) linear predictive coders,
5) noise-shaping coders

Lowpass filtering

Lowpass digital filtering is among the simplest methods. This method is shown in figure 1. A brief discussion of the method is made, without implementation details, in [3] and [24]. In some papers which discuss other methods the idea is considered as an introduction to or reference for more complicated, higher performance schemes [1],[8],[24]. One author notes that the lowpass method has been known for many years and has been used in voice-grade converters used in telephone systems [1]. Deyst has analyzed this method using an FIR filter [11]. However, very long filters having filter coefficients which require general multiplication are used in this analysis. Therefore, high-speed systems would be difficult to implement.

The method of Claasen et al.

The method of Claasen et al. [8] is shown in figure 2. In this method the signal and the quantization noise are conditioned before the lowpass filtering stage. The basic idea is to integrate the signal, quantize, then difference the samples. The differencing operation shapes the power spectrum of the noise so that most of the spectrum is out of the signal band. Integration is performed on the analog signal before quantization to minimize the effect of the differencing operation on the signal. Finally, the output is lowpass filtered to remove the out-of-band noise. The input

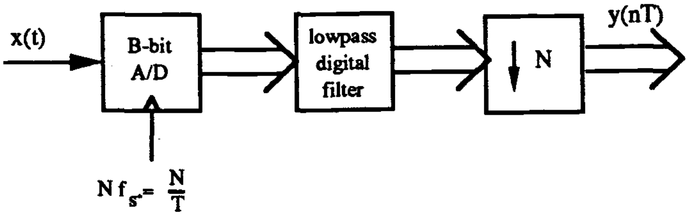
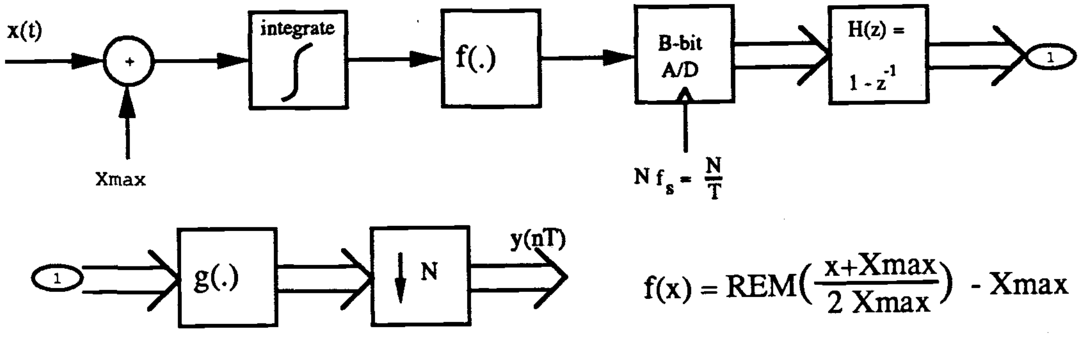


Figure 1 Lowpass filter method



$$f(x) = \mathrm{REM}\left(\frac{x+\mathrm{Xmax}}{2\ \mathrm{Xmax}}\right) - \mathrm{Xmax}$$

REM is remainder

$$g(x) = x - \mathrm{sgn}(x)$$

Figure 2 Claasen system

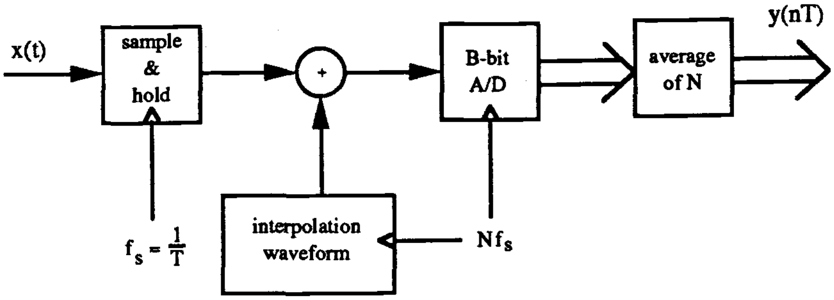


Figure 3 Belcher system.

signal is assumed to be limited to an amplitude range -Xmax < x(t) < Xmax. Adding Xmax to the input ensures that the signal which is integrated is non-negative. The non-linearity before quantization, f(.), is required so that the integrated signal does not saturate. The non-linearity after quantization, g(.), cancels this distortion. Because of the noise shaping, this method offers substantially better performance than the lowpass method--it has the best performance of the non-feedback methods. However, it is clearly more complex. It is not clear what advantage would be gained over noise shaping coders which use feedback, which offer similar performance for comparable complexity and have the advantage of greater noise shaping flexibility. This might explain why no expansions to the idea were found in the literature.

The method of Belcher

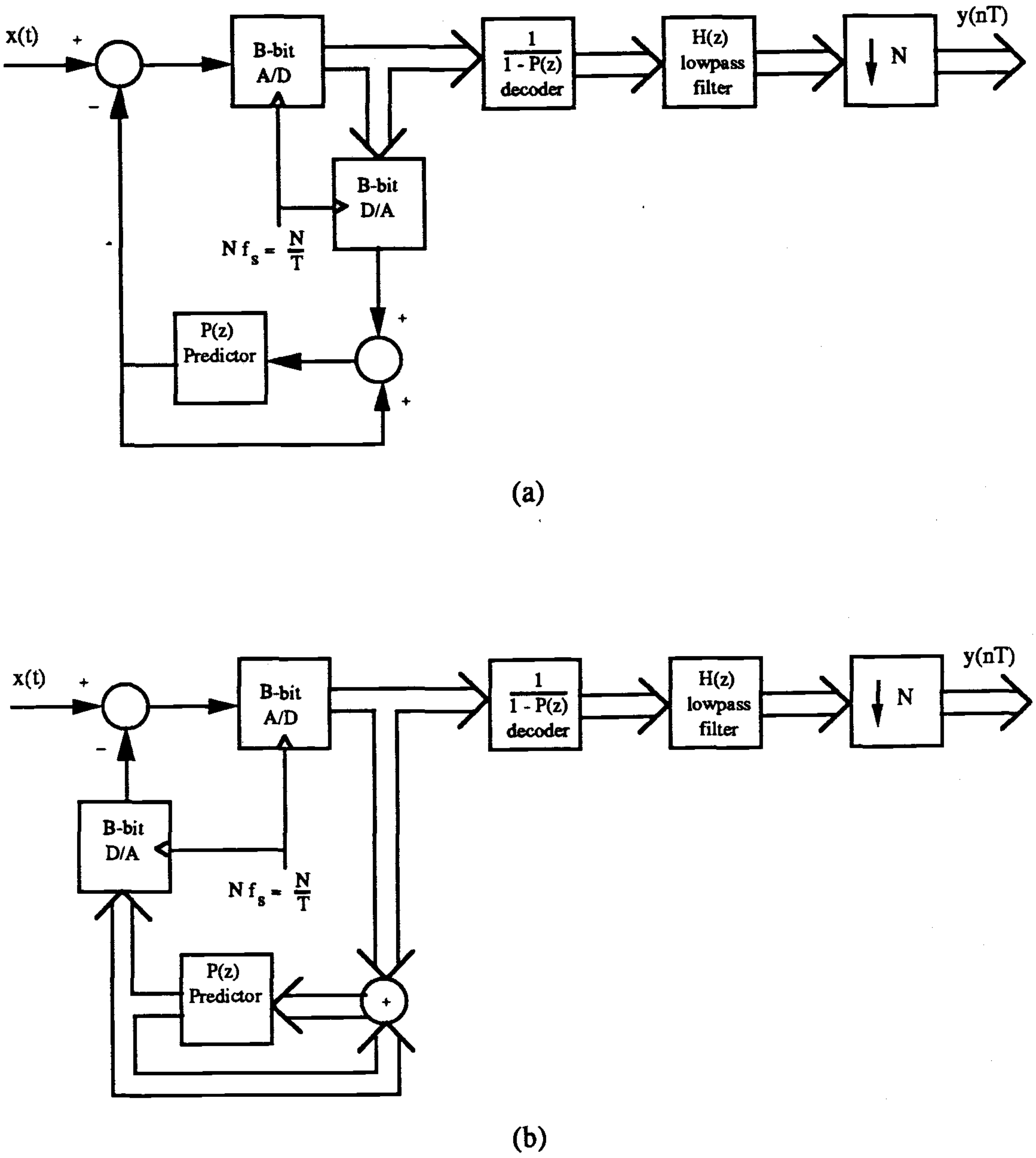
The system proposed by Belcher [4] is shown in figure 3. This method, along with the lowpass filtering method, is among the least complex schemes. Digital multiplication is not necessary if the $\frac{1}{N}$ factor implied in the averager is carried along as a scaling factor and is handled off-line. For $N=2^m$, scaling by $\frac{1}{N}$ could be be realized by an implied shift of the binary point by m places. This system has the effect of reducing the quantization uncertainty of the A/D converter by N. Deyst [11] has investigated the performance of this system and found that it is very sensitive to A/D converter non-linearities.

Linear predictive coders

The use of feedback in oversampled A/D converters dates back to the 1950s [10],[22]. First we consider systems termed linear predictive coders. This name comes from the telephony field where linear predictive coders have been used to reduce the bit rate for coded speech waveforms by removing redundancy (correlation) before quantizing [12]. Linear predictive coders can offer resolution enhancement without oversampling, but benefit significantly from oversampling [24]. This method is shown in figure 4. The transfer function in the feedback path can be analog or digital, depending on the location of the digital-to-analog converter (DAC). Popular methods, which fit into this category include delta modulation (DM), and differential pulse coded modulation (DPCM). Linear predictive coders operate by quantizing the difference between the input signal and a linear prediction of the current output signal. This difference signal has a smaller variance than the input signal and makes more efficient use of the dynamic range of the quantizer. The encoded output from the feedback loop is then "reconstructed" by adding the quantized difference to the linear prediction of the output.

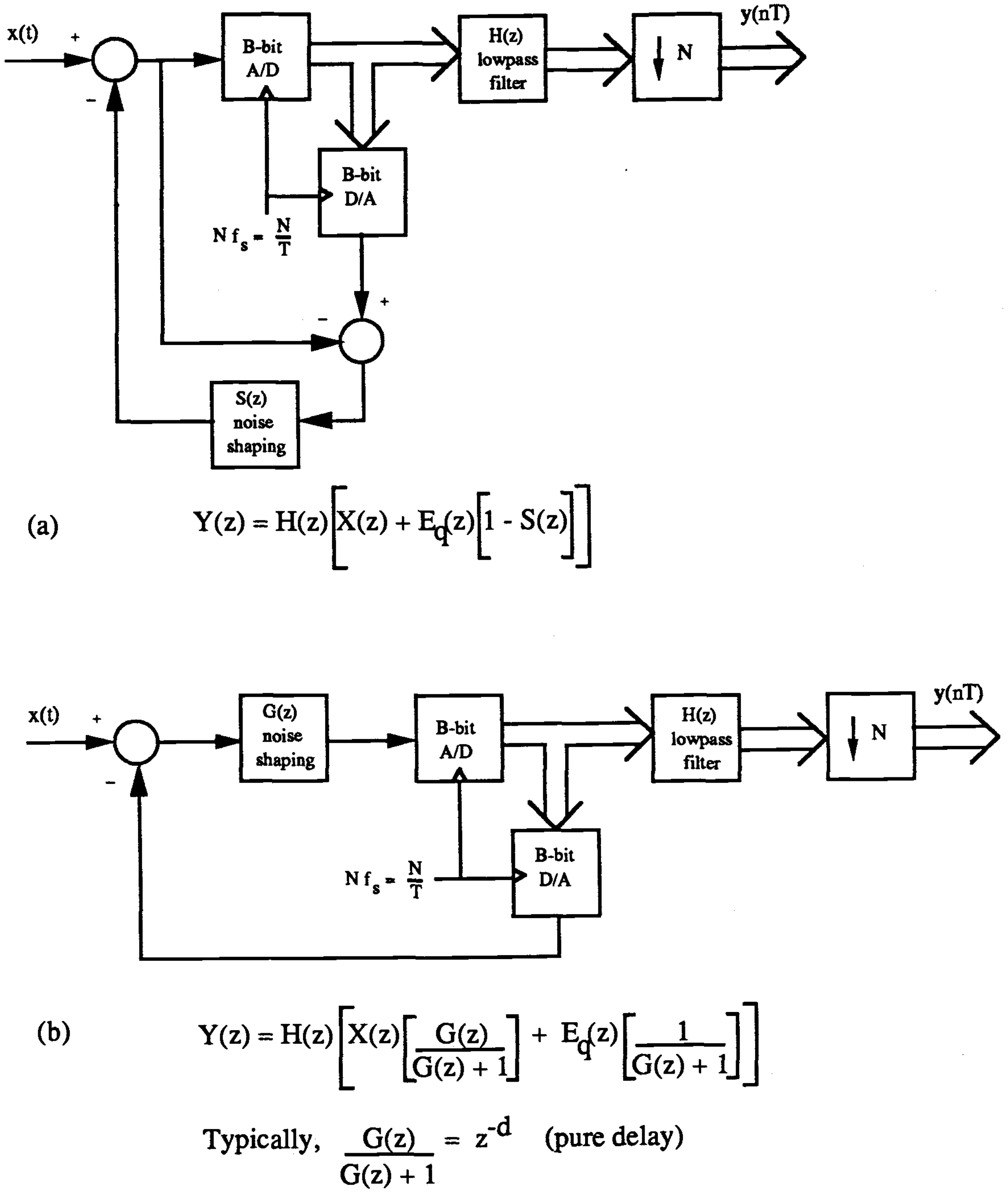
Noise shaping coders

The noise shaping coders are similar to the linear predictive coders. In figure 5, a general noise shaping coder is shown along with a popular noise shaping configuration, the delta-sigma modulator [14] -- also known as the sigma-delta modulator [1], [2], [16], [18] . The noise shaping coders get their name because they shape the power spectrum of the quantization noise such that little of the noise power remains in the signal band. Ideally, this is accomplished without affecting the signal. The net effect is almost the same as the Claasen et. al. system, but is accomplished using linear feedback. Theoretically, the linear predictive and noise shaping coders have been shown to offer the same resolution enhancement for a given order [24]. However, the noise shaping coders have the advantage of using a general lowpass decimator as the final stage, rather than a "reconstruction filter" which needs to match a given predictor filter in the feedback path.



$$Y(z) = H(z)\left[X(z) + E_q(z)\right]$$

Figure 4 Linear predictive coders. (a) Using a discrete-time analog predictor (commonly implemented using a switched capacitor circuit), and (b) using a digital predictor. The input output relation is the same for both systems and is given assuming a quantization noise source $E_q(z)$.



(a) $$Y(z) = H(z)\left[X(z) + E_q(z)\left[1 - S(z)\right]\right]$$

(b) $$Y(z) = H(z)\left[X(z)\left[\frac{G(z)}{G(z) + 1}\right] + E_q(z)\left[\frac{1}{G(z) + 1}\right]\right]$$

Typically, $\frac{G(z)}{G(z) + 1} = z^{-d}$ (pure delay)

Figure 5 Noise shaping coders. (a) General noise shaping coder, and (b) sigma-delta modulator coder. Input output relations are given assuming a quantization noise source $E_q(z)$. The discrete-time transfer functions S(z) and G(z) operate on analog signals. These are commonly implemented using switched capacitor circuits.

Comparison of methods

Figure 6 shows a comparison of the ideal performance of the different methods. In terms of A/D resolution enhancement in implemented systems, the noise shaping (specifically sigma-delta modulators) have offered the highest· performance to date [7], [16], [18], [25]. Usually sigma-delta modulators are constructed using a single-bit quantizer, which is inherently linear. This has been critical because the overall linearity is limited by the linearity of the DAC in the feedback path (for one-bit quantizers, the DAC is simply the quantizer output followed by a flip-flop). Recently, a method has been proposed to digitally correct for the nonlinearity of the feedback DAC. Using this method a 4-bit quantizer, in conjunction with oversampling by 128, has achieved simulated results of 21-bit resolution [7].

A primary problem with sigma-delta modulators for meeting the objectives proposed in this thesis is related to operational speed. Very fast filters which realize a z-domain transfer function (usually switched-capacitor) are required in the feedback loop. State-of-the-art switched capacitor filters using silicon technology offer sample rates of about 50 MHz [21], which is inadequate for the hundreds of MHz objective of this thesis. Another problem is that noise shaping causes an increase in total noise power [2]. The performance degradation due to poor lowpass filter stopband attenuation is more pronounced compared to the non noise-shaped lowpass method. Therefore, using a computationally simple filter capable of operating at high throughput rates can produce significantly poorer results than ideal.

To meet the objectives of this thesis, the lowpass method was chosen. A set of very simple digital filters are proposed which require no multiplication (not even shifting), only addition. These filters operate at the oversampled rate. A set of amplitude equalization filters operating at the decimated rate is also proposed to complete the method. Thus a two stage decimation filter is realized.

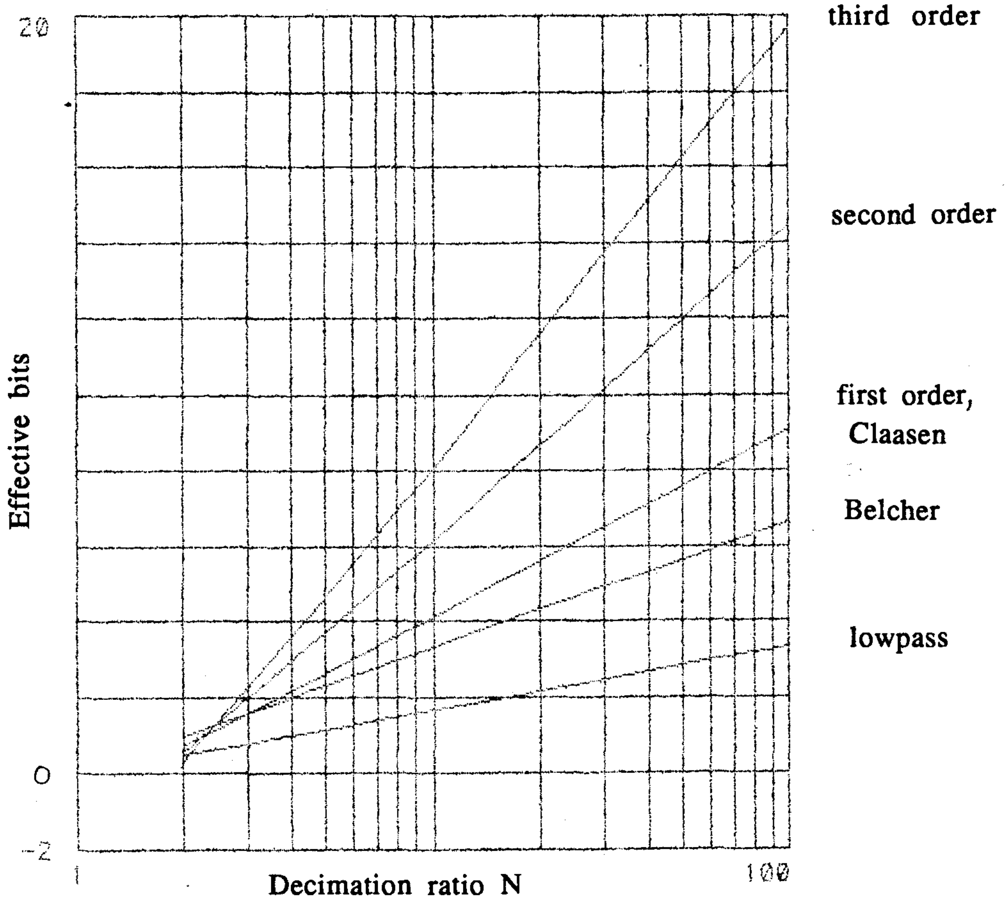


Figure 6

Ideal resolution enhancement in effective bits for different oversampling methods. The quantization error is assumed to be white noise. All lowpass filters used in methods are ideal. "N order" refers to an Nth order feedback coder (linear predictive or noise shaping).

1.3 Organization

This thesis examines the use of a particular family of computationally simple lowpass filters for enhancing A/D resolution. The performance of this method is dependent on the power spectrum of the quantization error. Chapter 2 presents an overview of models of the quantization noise found in the literature, as well as some comparisons of model predictions and simulated results. Chapter 3 introduces the family of filters and examines their characteristics. In chapter 4 a theoretical analysis is performed of the A/D resolution enhancement using this family of filters and a quantization noise model presented in chapter 2. Chapter 4 also examines the case of using noise shaping methods combined with a simple lowpass filter and compares this to the lowpass method proposed in this thesis. The validity of the theoretical analysis is demonstrated by simulations presented in chapter 5. Simulation results are also presented which examine the effectiveness of the family of filters in reducing alias errors caused by decimation. The validity of the beneficial effects of dither, which is discussed in chapter 2, is demonstrated by simulations. Chapter 6 presents the conclusions of the thesis. Finally, appendices are included which define errors and discuss error calculation methods used in conjunction with the simulations.

Chapter 2

Quantization Error Models

The lowpass method is very dependent on the broadband nature of the quantization error. This chapter examines common assumptions of the properties of quantization error and reviews some error models. Conclusions are made regarding the validity and use of these models.

The quantization error is defined by

$$e_q(nT) = Q[x(nT)] - x(nT) \qquad (1)$$

where T is the sampling period, x(nT) is the input signal, and Q[.] represents the uniform quantization operation. A graph of Q[x] is shown in figure 7a. A graph of the relationship of the quantization error as a function of input, is shown in figure 7b.

Usually, analysis of quantization error effects is made possible by assuming the error is a stochastic process which is statistically independent of the input signal. The non-linear quantizer can then be modeled as a linear system consisting of the input signal added to an independent noise source, as shown in figure 10b. Frequently, all the following assumptions are made to simplify analysis [19] :

1) $e_q(nT)$ is a stationary random process.
2) $e_q(nT)$ is uncorrelated with input sequence x(nT)
3) $e_q(nT)$ is an independent white noise process.
4) The probability density function of $e_q(nT)$ is uniform over the amplitude range of quantization error, q.

If the input signal is "complicated", such as speech or music, this model has been found to work well [19]. However, there are many cases where, clearly, this model performs poorly (eg. constant input, step function,

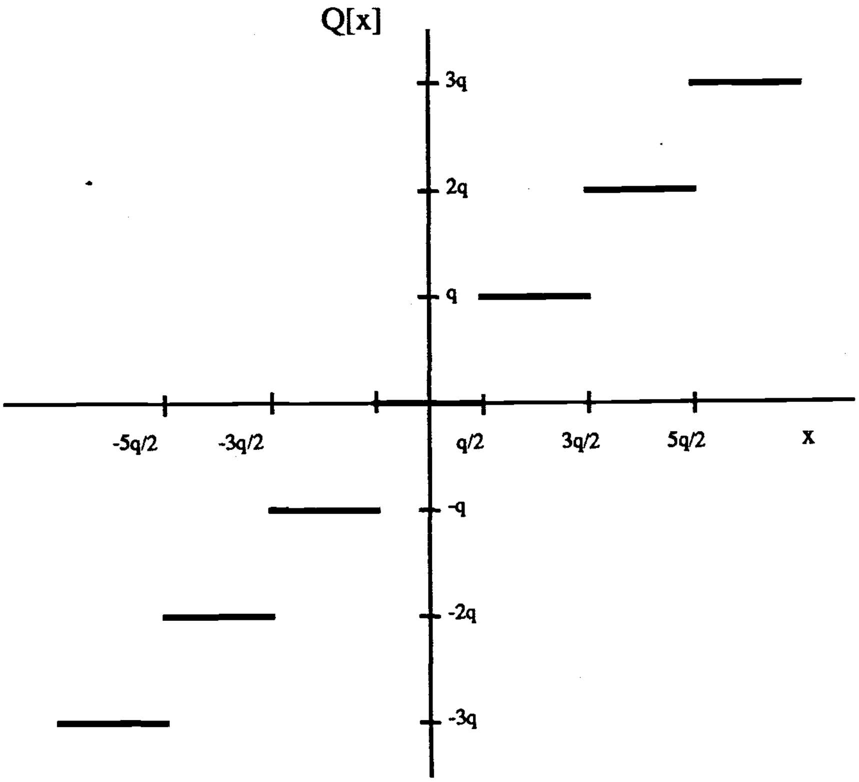


Figure 7a Uniform quantizer input-output relationship.

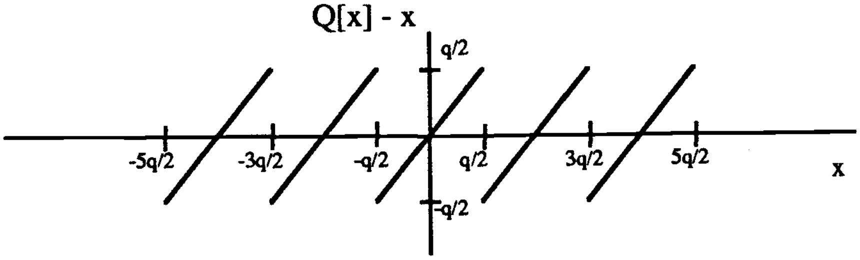


Figure 7b Uniform quantizer error as a function of input.

square wave). Mathematical models have been developed which are the basis for these assumptions and clarify their validity.

2.1 Models for stochastic inputs

First, models are presented which assume the input signal is stochastic. Widrow showed [26] that if

$$q < \frac{2\pi}{\xi_{max}} \tag{2}$$

then the quantization error can be shown to be an independent stochastic process (white noise) of uniform density and which is uncorrelated with the input,

$$E\{e_q(n)\ x(n+m)\} = 0 \tag{3a}$$

$$f_{e_q}(x) = \begin{cases} \frac{1}{q}, & \frac{-q}{2} < x < \frac{q}{2} \\ 0, & \text{otherwise} \end{cases} \tag{3b}$$

$$E\{e_q(n)\ e_q(n+m) = \begin{cases} \sigma_{e_q}{}^2 = \frac{q^2}{12}, & m=0 \\ 0, & \text{otherwise} \end{cases} \tag{3c}$$

$$E\{e_q(n)\} = 0 \tag{3d}$$

where q is the amplitude between quantization levels[1] and ξ_{max} is the highest significant (non-zero for an exact derivation) frequency in the characteristic function of the input process.[2]

[1] $q = (2^{-B}$ fullscale), where B is the number of bits in the quantizer, and fullscale is the analog range of the quantizer.

All these results were derived using a quantizer of infinite amplitude range. Infinite amplitude range is necessary for strictly meeting the bandlimited condition of the characteristic function. Sripad and Snyder have shown [23] that (2) is a sufficient, but not necessary for the model of (3) to hold. They derive a necessary and sufficient condition which is

$$\phi_{x_1 x_2}\left(\frac{2\pi l}{q}, \frac{2\pi k}{q}\right) = 0 \qquad \forall \;\; l \neq 0 \text{ and } k \neq 0 \qquad (4)$$

where $\phi_{x_1 x_2}(\zeta_1, \zeta_2)$ is the second order characteristic function of the sampled input x(nT) ($x_1 = x(nT)$ $x_2 = x(mT)$).

The significance of these result can be seen for the example of a Gaussian input. Using the results of an analytical model related to (4), it has been shown that for a first order Gaussian input the uniform noise model is a very close approximation for $\frac{\sigma}{q} \geq 0.7$, where σ is the standard deviation of the input [23]. For the case of a second order Gaussian input, it has been shown that the quantization noise is practically uncorrelated for $q = \sigma$ and $\frac{\sigma_{12}}{\sigma^2} < 0.9$ where σ_{12} is the 1,1 moment of the input [26]. Using these results,

[2] If the input process is n-th order, then $\xi_{max} = \sup\{\xi_{max_i}\}$, $i = 1,\ldots,n$. For example, if the input process is second order the characteristic function is

$$\phi_{x_1 x_2}(\xi 1,\xi 2) = \int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} f_{x_1 x_2}(x1,x2)\, e^{j(x1\xi 1 + x2\xi 2)}\, dx_1\, dx_2$$

where $x_1 = x(n)$ and $x_2 = x(m)$. ξ_{max_1} is the highest $\xi 1$ for which $\phi_{x_1 x_2}(\xi 1,\xi 2)$ is non-zero and ξ_{max_2} is the highest $\xi 2$ for which $\phi_{x_1 x_2}(\xi 1,\xi 2)$ is non-zero. And

$$q < \frac{2\pi}{\sup\{\xi_{max_1},\xi_{max_2}\}}$$

for the model to hold.

it has been qualitatively stated that if the dynamic range of a variable being quantized extends over several quantization levels, q, the model in (3) holds [26]. However, it is easy to think of some deterministic inputs covering a wide dynamic range for which this model breaks down -- basically any signal that varies less than q for long periods of time (eg. a square wave).

2.2 Models for deterministic inputs

In response to this problem, a model has been proposed based on the amplitude distribution function (ADF) of the derivative of a deterministic signal [9]. The model of the power spectrum of the quantization noise is

$$S_{e_q}(\Omega) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{p_{\dot{x}}(\Omega/2\pi k)}{k^3} \qquad (5)$$

where $p_{\dot{x}}(a)$ is the amplitude distribution function of the derivative of the input signal[3] , and Ω is is the continuous-time frequency variable. For the usual case of discrete-time the expression is given as,

$$S^d_{e_q}(e^{j\omega}) = \frac{1}{T^2} \sum_{m=-\infty}^{\infty} S_{e_q}(\Omega + m\frac{2\pi}{T}) \qquad (6)$$

where T is the sample period.

For the important case of a sinusoid $x(t) = A\sin(\Omega_0 t)$, the ADF is,

[3] All references to amplitude used in this model are normalized to the amplitude of the quantization step, q.

$$p_x(a,A) = \begin{cases} \frac{1}{A}\frac{1}{\sqrt{1-(a/A)^2}}, & |a| \le A \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

and the ADF of the derivative of x is $p_{\dot{x}}(a) = p_x(a,\Omega_0 A)$. Inserting (7) into (5) and then (5) into (6) we can find the model of the quantization noise for the sinusoid sampled with period T. The authors show that if

$$T >> \frac{1}{\Omega_0 A} \qquad (8)$$

the noise power spectrum is flat. A comparison of the noise power spectrum predicted by this model and simulated results is shown in figures 8a - 8c. Note that the model performs best for the case of the smallest oversampling ratio. Also note that for the higher oversampling ratios, for N=500 and N=5000, the noise is becoming significantly correlated. For the former, the noise is highpass, and for the later it is lowpass.

2.3 Effect of dither on quantization noise

A small dither signal can be added to the input signal before quantizing to whiten the power spectrum of the quantization noise and decorrelate the signal from the quantization noise. The dither signal can be subtracted from the samples after quantization. In an oversampled acquisition system, most of the broadband dither power can be removed by lowpass filtering [6].

A necessary and sufficient condition for a dither signal, w(t), which will make the quantization noise white and uncorrelated with the input is [20],

$$\phi_W\left(\frac{2\pi k}{q}\right) = 0 \quad \forall \quad k \neq 0 \qquad (9)$$

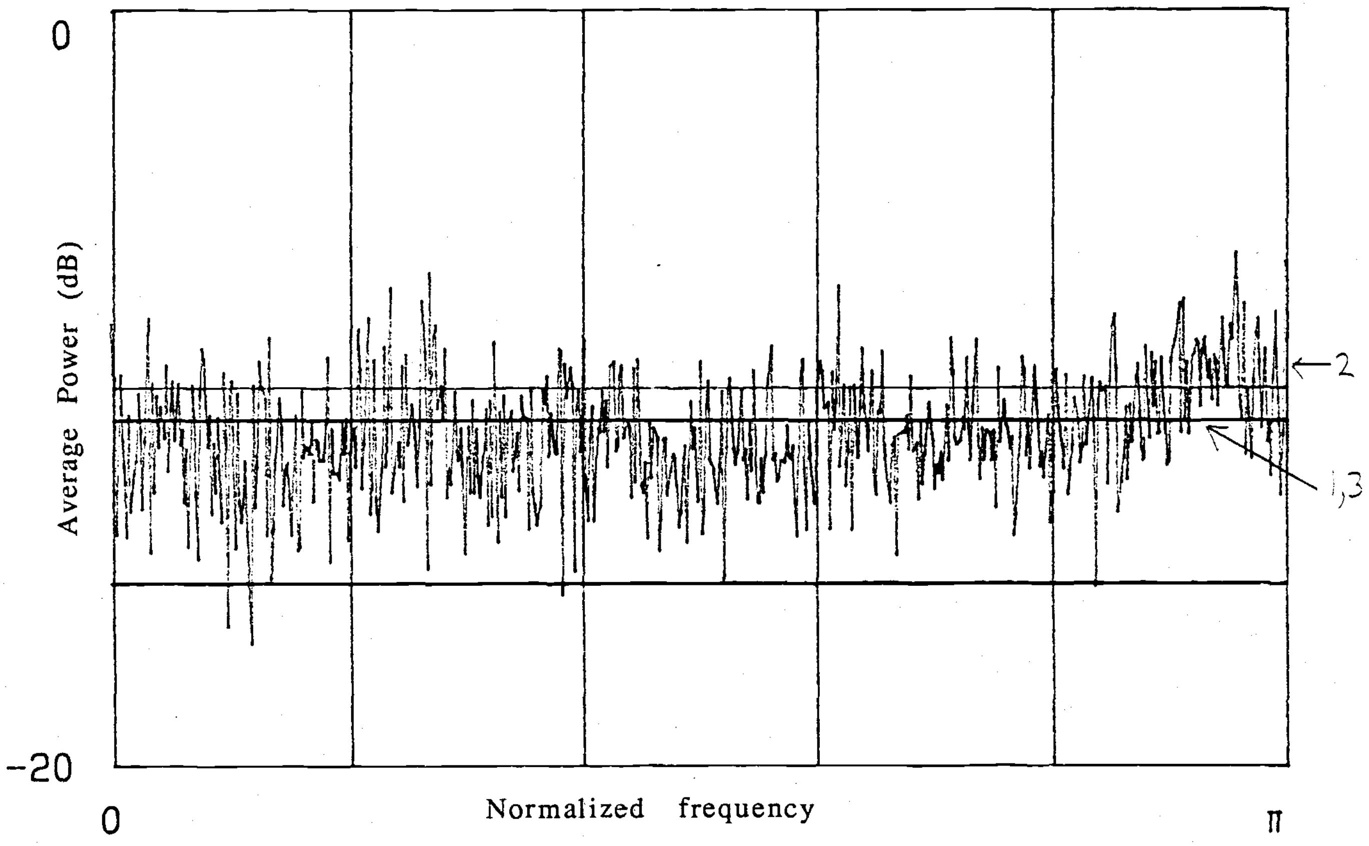


Figure 8a

Power spectrum of quantization noise for oversampling ratio of 50, where the input is a sinusoid covering 244 quantization levels: 1) deterministic model, 2) simulated results of average of 10 periodograms of slightly different input frequency, and 3) mean square value over frequency for simulated results.

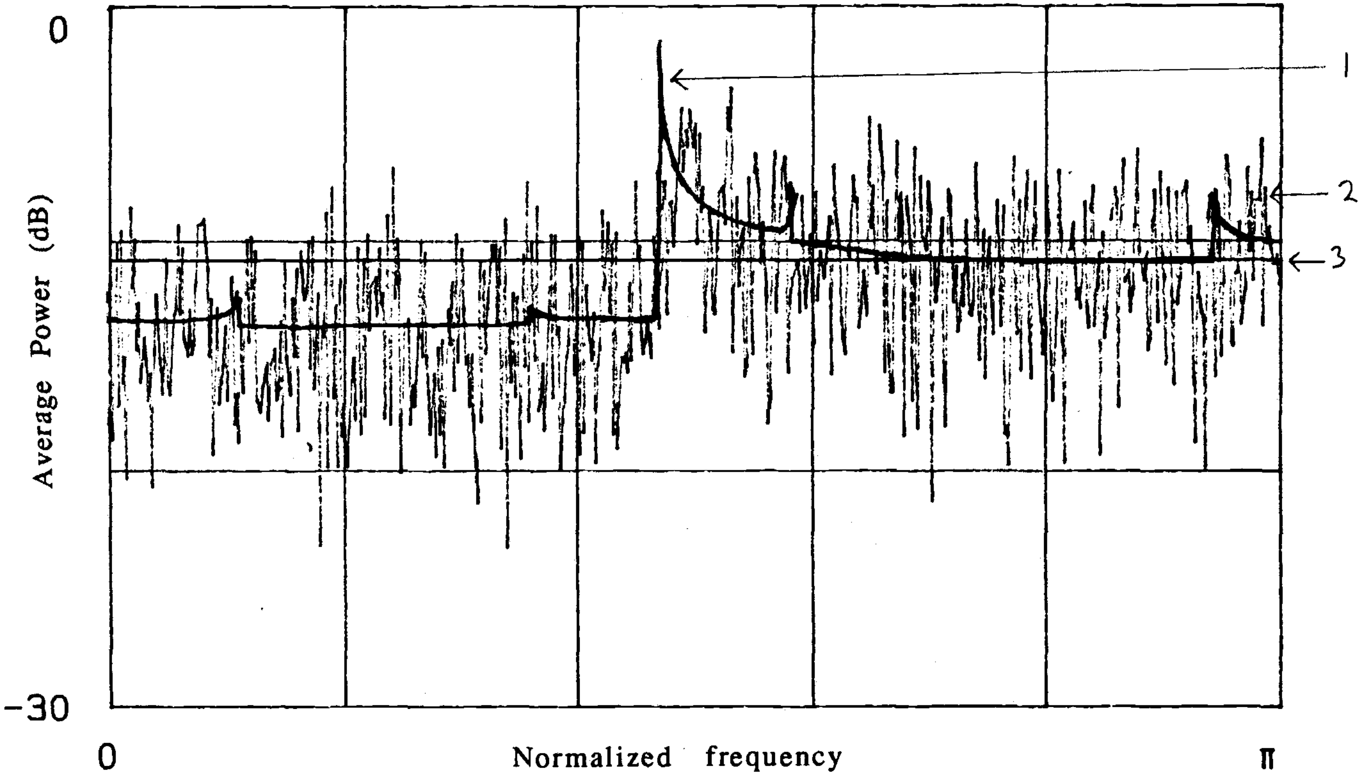


Figure 8b

Power spectrum of quantization noise for oversampling ratio of 500, where the input is a sinusoid covering 244 quantization levels: 1) deterministic model, 2) simulated results of average of 10 periodograms of slightly different input frequency, and 3) mean square value over frequency for simulated results.

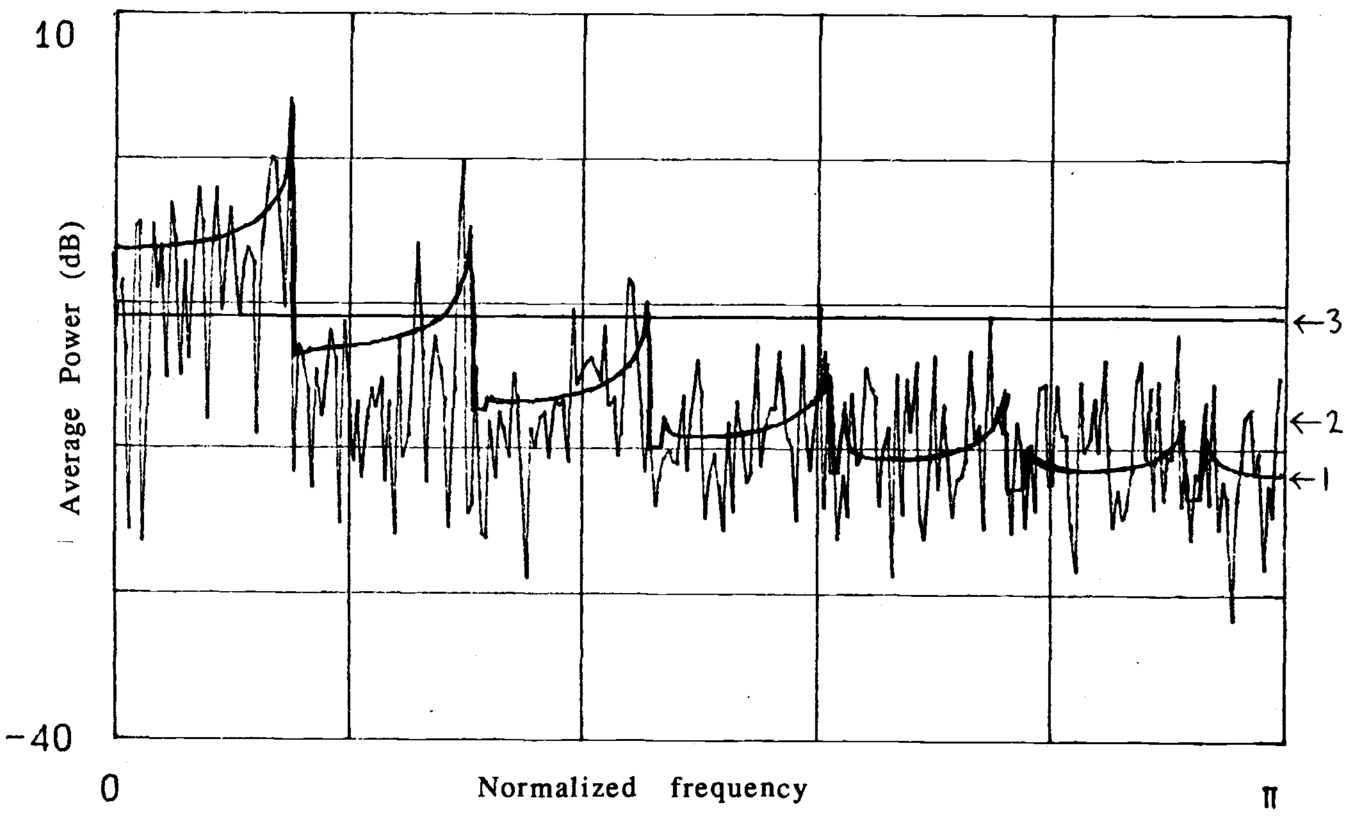


Figure 8c

Power spectrum of quantization noise for oversampling ratio of 5000, where the input is a sinusoid covering 244 quantization levels: 1) deterministic model, 2) simulated results of average of 10 periodograms of slightly different input frequency, and 3) mean square value over frequency for simulated results.

Some common examples of the probability density function which satisfy this condition are shown in figure 9.

Using the results from the stochastic models it can be stated that if the signal has random amplitude fluctuations that cover a few quantization intervals, then the white noise model of (3) should be applicable. For signals which do not fluctuate randomly, the white noise model is not necessarily a good model, even if the signal amplitude covers many quantization intervals. This was demonstrated in figures 8b and 8c for the case of highly oversampled sinusoids. The deterministic model for the quantization power spectrum could be used for predicting results. In an actual acquisition system it is reasonable to suspect that there will be small additive noise which are uncorrelated with the input -- or a dither signal could be purposely added. If the dither satisfies (9), it will cause the quantization noise to be white. Even if it does not satisfy (9) exactly, it will tend to decorrelate the quantization noise. Therefore, for an actual acquisition system, the white noise model should be reasonably accurate.

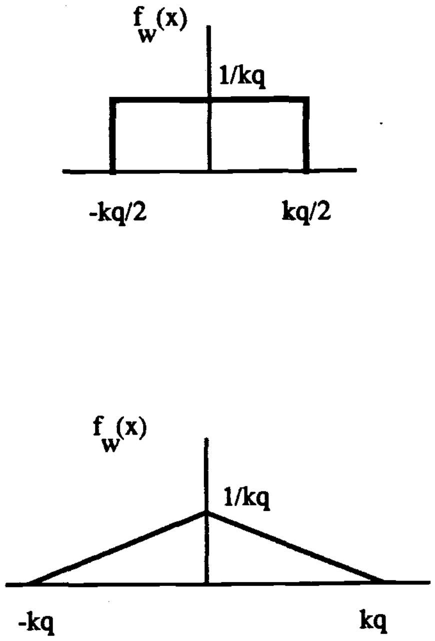


Figure 9 Examples of dither probability density functions which cause the quantization noise to be white.

Chapter 3

Oversampled Acquisition System using a Simple Averaging Lowpass Filter

The lowpass method using a family of very simple finite impulse response (FIR) lowpass filters (known as comb filters or averaging filters) is proposed for reducing quantization error. This family will be referred to collectively as the MA (moving average) filter. These filters will operate at the oversampled rate. The only arithmetic operation they require is addition. A complimentary family of amplitude equalization filters, which operate at the decimated rate, is also proposed. These filters will be referred to collectively as the inverse filter. The inverse filter requires general multiplication. The MA filter and the inverse filter form a two-stage decimator. The architecture of the complete acquisition system is shown in figure 10a. Figure 10b gives a more general system diagram which will be referred to throughout the text.

3.1 MA Filter characteristics

The impulse response of the MA filter is

$$h_{MA}(n) = \begin{cases} \frac{1}{N}, & 0 \le n < N \\ 0, & \text{otherwise} \end{cases} \qquad (10)$$

where the $\frac{1}{N}$ scaling factor can be eliminated from the filter implementation and computed off-line, if necessary. This filter requires only the addition operation and multiplication is not necessary. Another advantage of this filter is that the width of the passband of the filter is

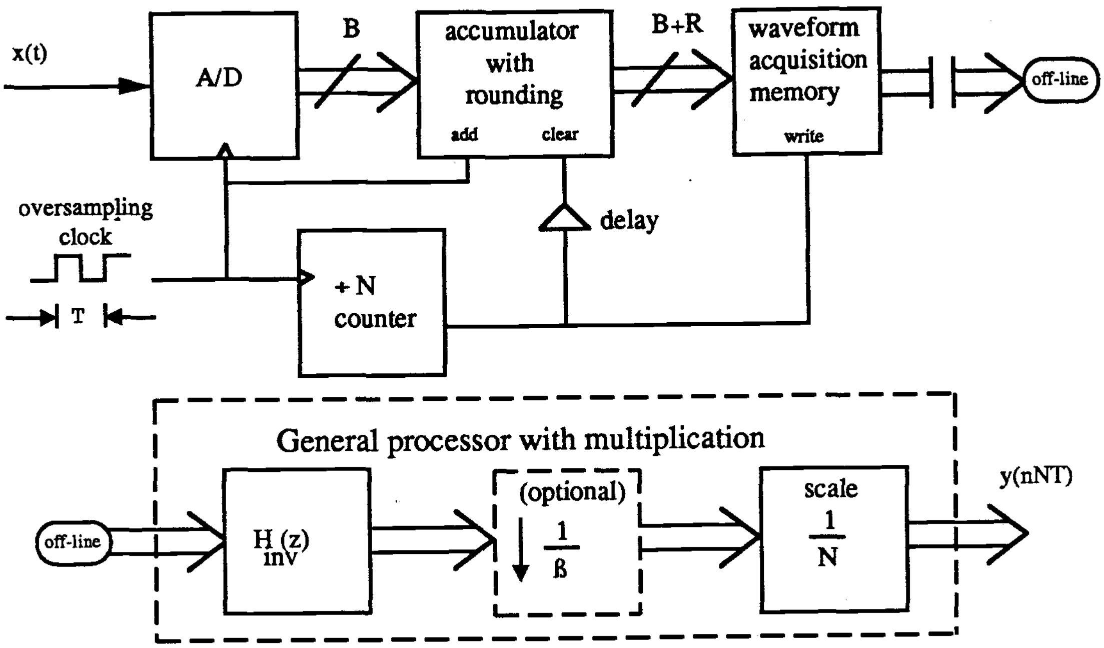


Figure 10a System architecture of lowpass system using MA filter

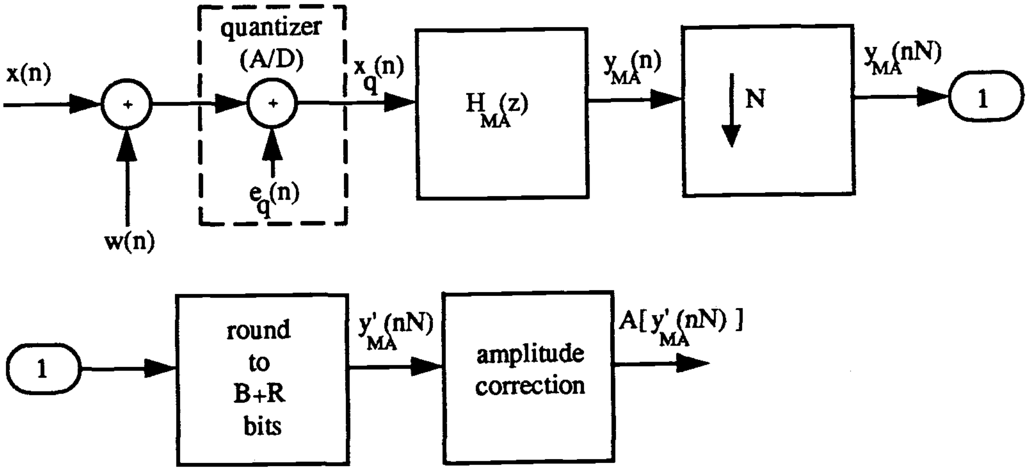


Figure 10b General bock diagram of lowpass system using MA filter

inversely proportional to the filter length, N. Therefore, to realize the filtering and decimation process, it is possible to accumulate N inputs and store them directly to waveform memory. If this relationship were not true, additional buffering would be necessary.

The frequency response of the filter is,

$$H_{MA}(z) = \frac{1}{N}\sum_{i=0}^{N-1} z^{-i} \qquad (11)$$

Evaluating at $z=e^{j\left(\frac{\omega}{f_s}\right)}$,

$$HMA(e^{j\omega}) = e^{-j\left(\frac{\omega(N-1)}{f_s}\right)} \frac{\sin\left(\frac{\omega N}{2f_s}\right)}{N\sin\left(\frac{\omega}{2f_s}\right)} \qquad (12)$$

,where f_s is the sampling frequency. The frequency response is shown in figure 11 for N=10, and figure 12 for N=50. If the signal is decimated by N after filtering, signal components will fold about the frequency $2\pi f_s/2N$, (normalized frequency of π/N). Therefore the passband will be defined by,

$$-\frac{\pi f_s}{N} < \omega < \frac{\pi f_s}{N} \qquad (13)$$

The nulls of the filter are at $\omega_{null} = \frac{2\pi n f_s}{N}$, n=1,...,N-1. An interesting consideration is that, although the filter response has high side lobes, if the desired passband is much smaller than π/N, the nulls in the response will reject the bands which will fold into the passband of the decimator.

Figures 13-15 summarize the magnitude response characteristics of the MA filter. The maximum passband error for various values of N is shown in figure 13 where the error is the difference between the ideal LPF and the MA filter at the cutoff frequency of $\omega_c = \frac{\pi f_s}{N}$, that is,

$$\max(ep) = 1 - \left|HMA\left(e^{j\frac{\pi fs}{N}}\right)\right| \qquad (14)$$

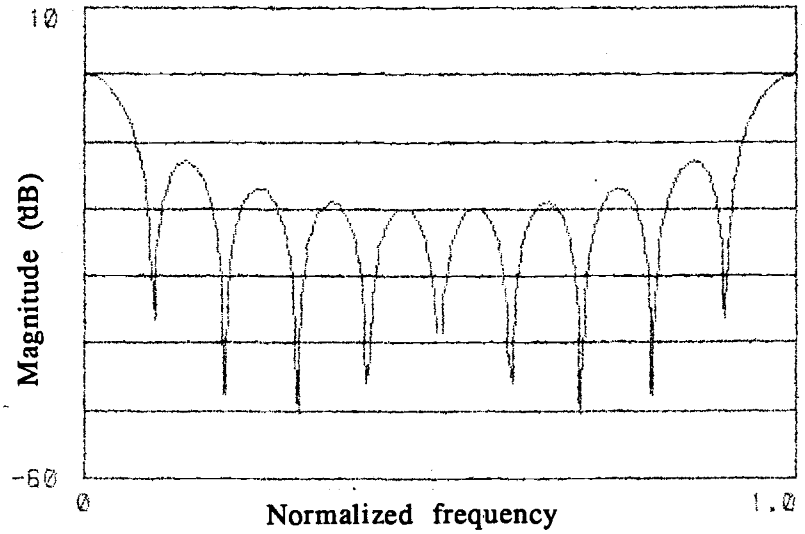


Figure 11 Magnitude of the MA filter transfer function for N = 10.

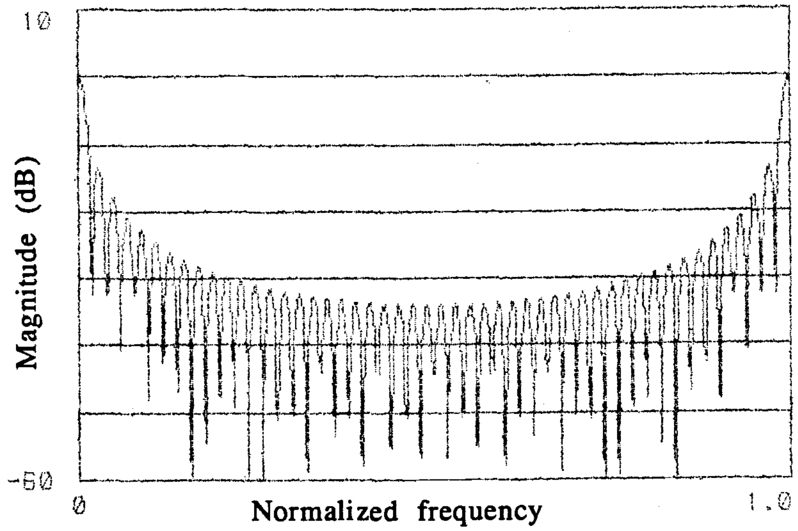


Figure 12 Magnitude of the MA filter transfer function for N = 50.

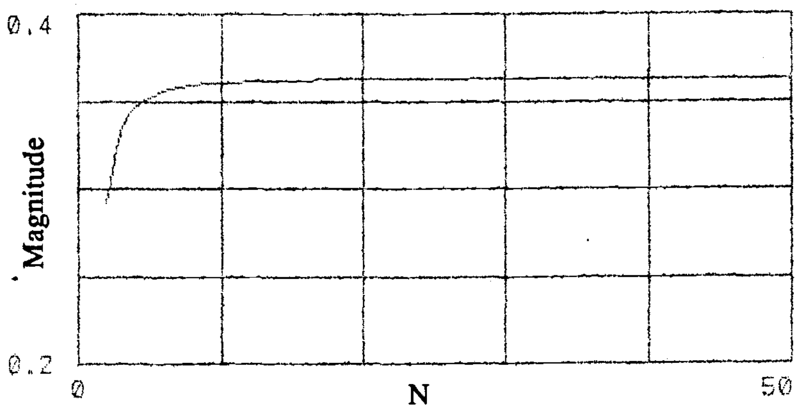


Figure 13 Maximum passband error for the MA filter as a function of filter length, N.

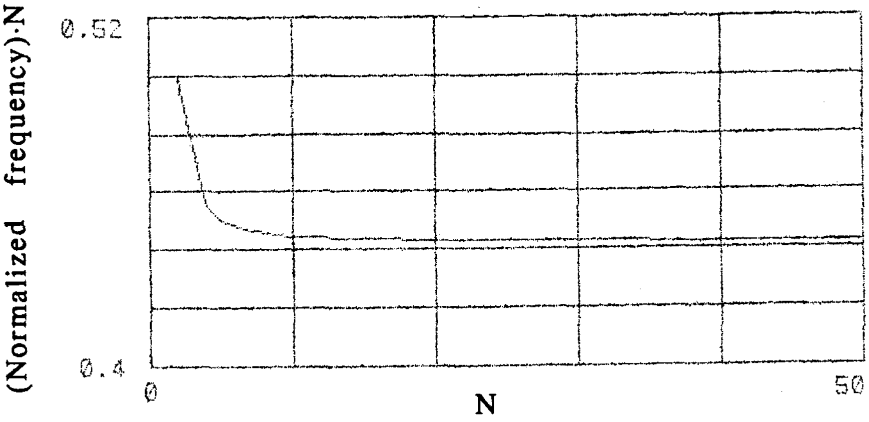


Figure 14 Half-power frequency (-3dB) for the MA filter as a function of filter length, N. Frequency is multiplied by N to adjust it to the filter passband.

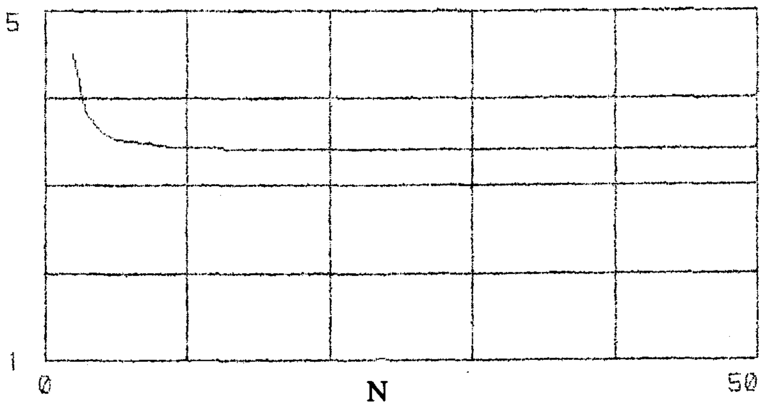


Figure 15 Band energy ratio for the MA filter as a function of filter length, N.

The error reduces to about 0.37 for N ≥ 10. Figure 14 depicts the half-power (-3 dB) frequency as a function of N (here, frequency is multiplied by N to scale to filter passband). Note that this does not significantly change for N ≥ 10. Finally, figure 15 shows the ratio of passband energy of the filter over the total energy in all bands except the passband of the filter, that is,

$$\text{band energy ratio} = \left(\frac{\text{passband energy}}{\text{otherband energy}}\right) \tag{15}$$

where[4] ,

$$\text{passband energy} = \frac{1}{2\pi}\int_{\frac{-\pi}{N}}^{\frac{\pi}{N}} |H_{MA}(e^{j\omega})|^2 \, d\omega \tag{16}$$

$$\begin{aligned}\text{otherband energy} &= \text{total energy} - \text{passband energy}\\ &= \frac{1}{N} - \text{passband energy}\end{aligned} \tag{17}$$

This ratio is a measure of filter ideality with respect to removing out-of-band noise.

In observing these graphs it can be seen that the performance of the MA filter does not improve significantly by increasing N beyond 10, except that the passband becomes more narrow with N. Therefore, we do not

[4] See [19] p. 572 for:

$$\text{total energy} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_{MA}(e^{j\omega})|^2 \, d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{\sin\left(\frac{\pi f N}{f_s}\right)}{N\sin\left(\frac{\pi f}{f_s}\right)}\right)^2 d\omega = \frac{1}{N}$$

expect the resolution enhancement to become more like the ideal lowpass filter case with increasing N.

3.2 Amplitude equalization methods and the inverse filter

The passband amplitude errors caused by the MA filter could be compensated by various methods.

If the input signal is known to contain a single frequency (ie. a sinusoid), the output could be scaled by $1/|H_{MA}(e^{j\omega_0})|$, where ω_0 is the frequency of the input. For the general case, an inverse filter could be applied to the decimated output. This filter has a frequency response of,

$$|H_{inv}(e^{j\omega})| = \frac{1}{|H_{MA}(e^{j\frac{\omega}{N}})|} \qquad , |\omega| < \pi \qquad (18)$$

The frequency response of this filter for N = 5 is shown in figure 16. The inverse filter increases the signal amplitude after decimation to cancel the undesirable non-flat passband of the MA filter. Note that by correcting (increasing) the signal amplitude, the acquisition errors (eg.. quantization, aliasing, other system noise) will be increased.

An alternative scheme which trades waveform memory for performance can be devised. Suppose that an input signal is bandlimited to only a fraction of the MA filter passband ($|\omega| < \omega_c$) or it is desired to only preserve this band of the signal. In this case it is only required to correct the amplitude over the fraction of the passband. From figure 16 it can be seen that the magnitude of the inverse filter frequency response is monotonically increasing with $|\omega|$ ($|\omega| < \pi$). Therefore, the increase in acquisition error caused by amplitude correction can be decreased if we only correct the amplitude over this fraction of the MA filter passband. To extend this idea, we can design the higher frequency band of the amplitude correction filter to approximate zero to reduce the acquisition errors in this frequency band, followed by a decimation by $\frac{1}{\beta}$, where β is the fraction of the MA filter passband which the signal covers. This scheme can be seen as oversampling by a higher ratio. However, it has two advantages compared to simply oversampling by a higher ratio using H_{inv} as defined in (18). One is that the acquisition errors are increased less by the process of amplitude correction. The second is that if there are large, undesirable signal components just beyond the filter passband, $|\omega| < \omega_c$, the alias errors will be reduced (see figure 17). This method requires more waveform memory because the final decimation by $\frac{1}{\beta}$ is performed using samples that have already been stored in the waveform memory.

The inverse filter and the alternate inverse filter are ideal filters. A Linear phase FIR approximation to the ideal inverse filter is shown in figure 18 for N=5. This filter was designed using the Parks et. al program [17] with the desired magnitude subroutine modified appropriately.[5]

There are three main conclusions from this chapter. The first is that the MA filter performs well at removing the quantization noise. In fact, since the total energy of the MA filter is $\frac{1}{N}$, which is the same as an ideal lowpass filter, the MA filter is as effective as an ideal lowpass filter in reducing the variance of white noise. The second conclusion is that the characteristics of the MA filter do not become significantly closer to the lowpass filter of the same cutoff frequency by increasing the filter length. Finally, a primary disadvantage of the MA filter is its non-flat passband response. This can be corrected at the expense of increasing the noise. For the general case of the signal covering the entire passband of the filter, the benefit of compensating for the non-flat passband response outweighs the negative effect of increasing the noise, since the passband error has a larger amplitude than the noise.

[5] If $|H_{inv}(e^{j\omega})|$ is used as the desired magnitude, the error is not equiripple--there is a negative bias to the error which increses with frequency. In an attempt to reduce this bias, the desired magnitude was altered slightly from $|H_{inv}(e^{j\omega})|$.

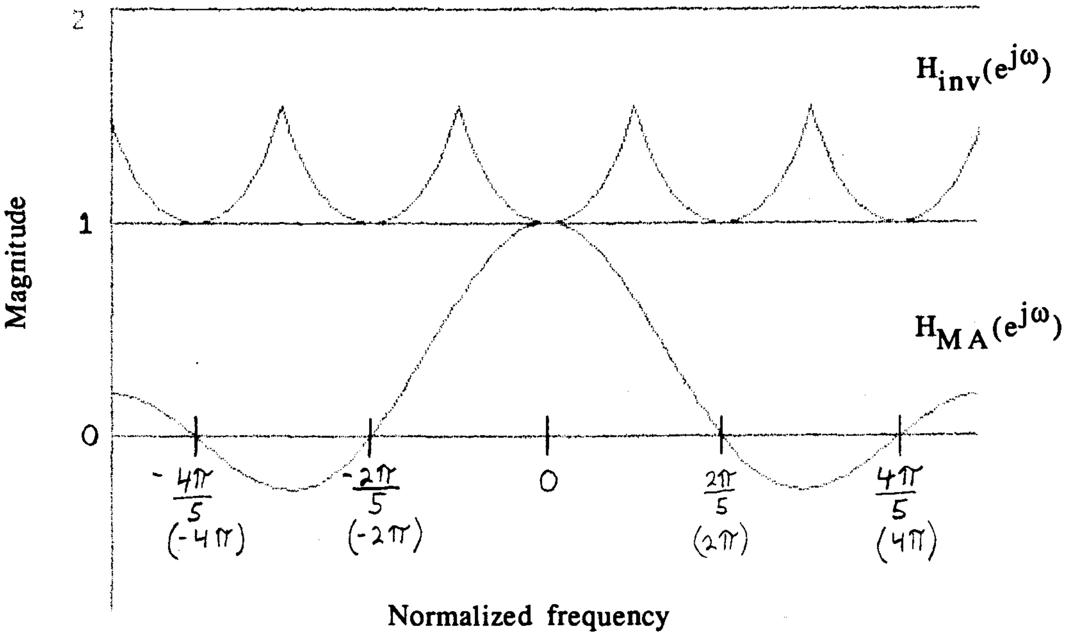


Figure 16 Frequency response of inverse filter and corresponding MA filter (N=5). The frequency scale in parentheses is for the inverse filter.

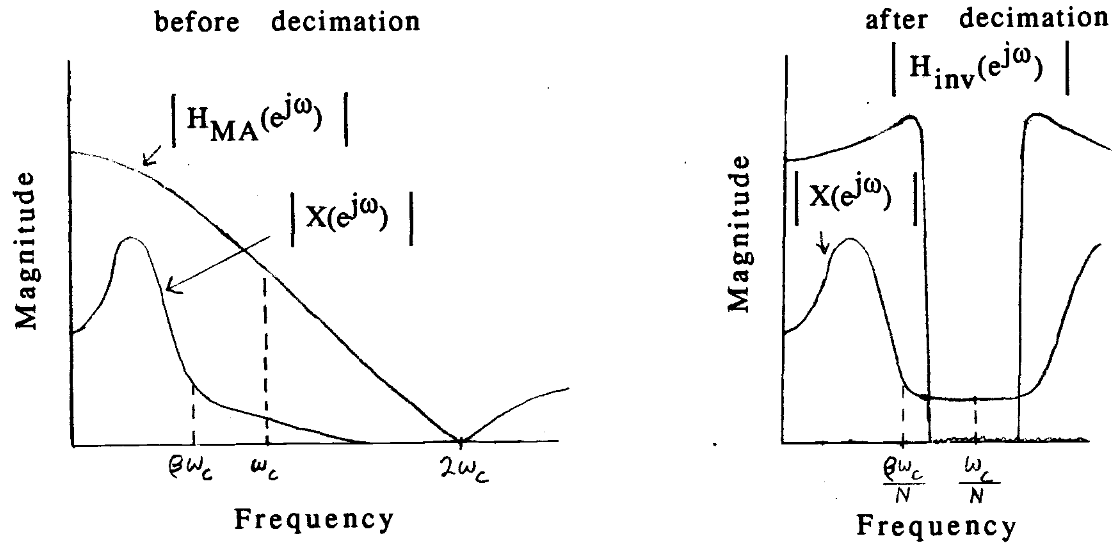
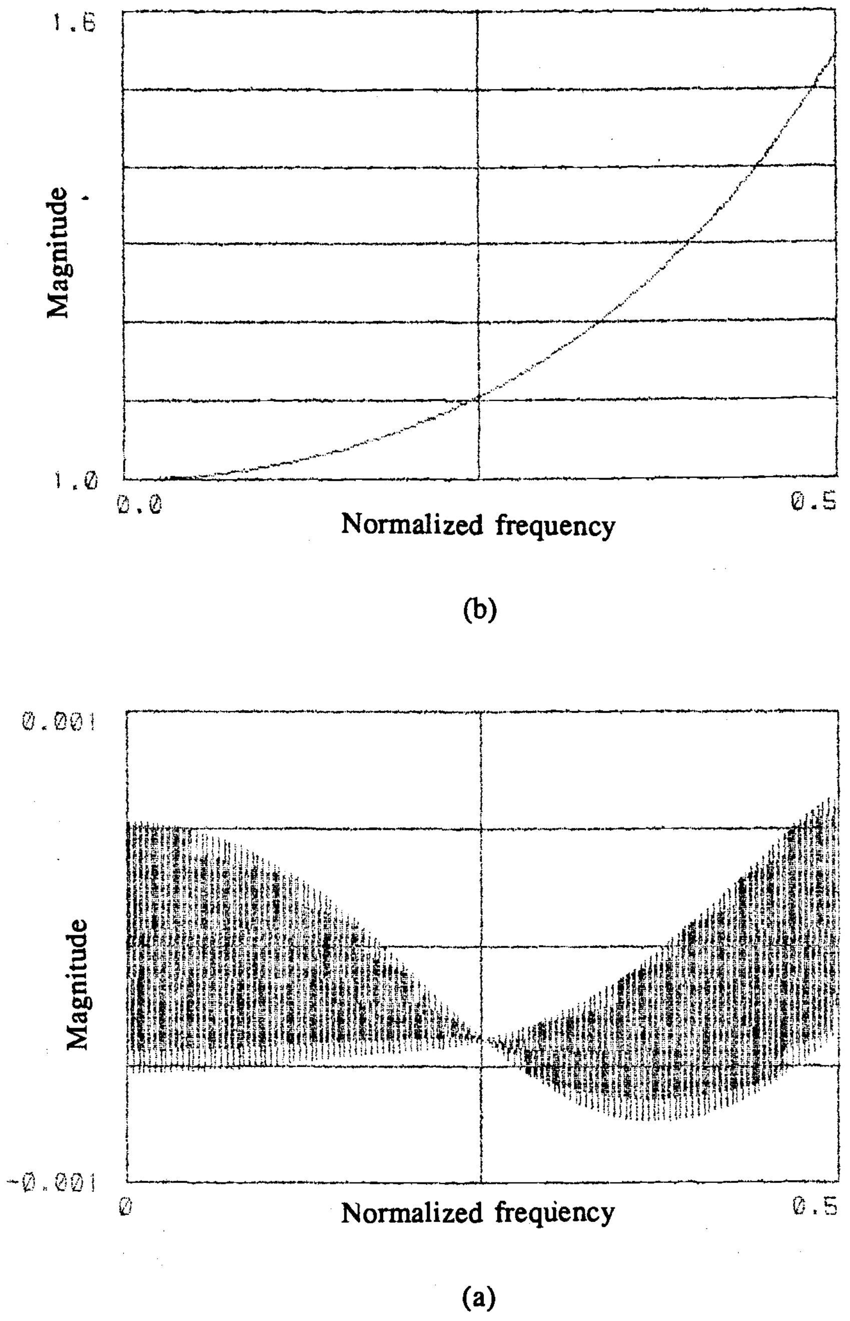


Figure 17 An alternate inverse filtering scheme used when the desired input signal band is a fraction (b) of the MA filter passband. After decimation, the undesirable signal beyond ω_c is aliased, but is eliminated by the stopband of the alternate inverse filter. Other acquisition errors in this band are also eliminated.



(b)

(a)

Figure 18 Example FIR realization of the inverse filter for N=5. The filter has 511 floating-point coefficients. (a) Magnitude response, and (b) error of magnitude from ideal.

Chapter 4

Theoretical System Performance

In this chapter, the performance of the MA lowpass system for increasing resolution of the A/D converter is considered. First effective bits is introduced as a measure of A/D converter resolution. Then an analysis of the resolution is given assuming the white noise model for the quantization noise. An interesting comparison is then made between the performance of the lowpass MA system alone compared to that of using the MA filter with noise shaping coders. Finally, the worst case performance is analyzed.

4.1 Effective bits

Effective bits has been defined as an empirical method for calculating the average resolution of an A/D converter using sinusoidal inputs [13]. In this thesis the measure will be used in a broader sense. The general usage will be explained in context. For more specific information, see appendix 2.

4.2 Theoretical analysis of improvement of SNR and effective bits

No aliasing effects caused by the decimation stage are considered here. Because of the signal dependency of alias errors, these effects will be examined through simulations.

The quantization noise in different parts of the oversampled system is labeled in figure 19. Assume, for the moment, that $e_q'(n)$ is decimated by factor, M, which is different than the length of the impulse response of the MA filter, N. The autocorrelation of the noise $e_q'(nM)$ is,

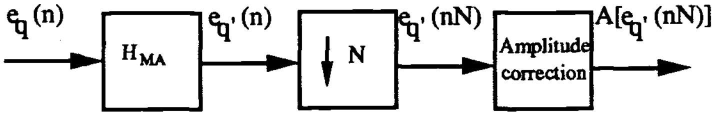


Figure 19 Quantization error notation at different stages in the system.

$$R_{e_{q'}}(kM) = E\{eq'(nM)\ eq'((n+k)M)\}$$

$$=E\{\sum_{i=-\infty}^{\infty} h_{MA}(nM-i)\ e_q(i)\ \sum_{j=-\infty}^{\infty} h_{MA}((n+k)M-j)\ e_q(j)\}$$

$$= \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h_{MA}(nM-i)\ h_{MA}((nM-j+kM))\ R_{e_q}(i-j) \qquad (19)$$

where it is assumed that $e_q(n)$ is wide-sense stationary,

$$R_{e_q}(i-j) = \sigma_{e_q}^{\ 2} \cdot \delta(i-j) \qquad (20)$$

Substituting (20) into (19) results in,

$$R_{e_{q'}}(kM) = \sigma_{e_q}^{\ 2} \sum_{i=-\infty}^{\infty} h_{MA}(nM-i)\ h_{MA}((nM-i+kM)) \qquad (21)$$

$$R_{e_{q'}}(kM) = \sigma_{e_q}^{\ 2} \sum_{p=-\infty}^{\infty} h_{MA}(p)\ h_{MA}((p+kM)) \qquad (22)$$

Now consider the case that the decimation factor, M, is greater than or equal to the MA filter length, N (in the proposed system M = N). In this case, (22) can be reduced to,

$$R_{e_{q'}}(kM) = \sigma_{e_q}^{\ 2} \sum_{p=-\infty}^{\infty} h_{MA}^{\ 2}(p)\ \delta(kM) \qquad (23)$$

The power spectrum of the error is then,

$$S_{e_{q'}}(z) = \sum_{k=-\infty}^{\infty} \left(\sigma_{e_q}^{\ 2} \sum_{p=-\infty}^{\infty} h_{MA}^{\ 2}(p)\ \delta(kM) \right) z^{-k}$$

$$= \sigma_{e_q}{}^2 \sum_{p=-\infty}^{\infty} h_{MA}{}^2(p) = \sigma_{e_q}{}^2 \sum_{p=0}^{N-1} \left(\frac{1}{N}\right)^2 = \frac{\sigma_{e_q}{}^2}{N} \tag{24}$$

It is seen that the power spectrum of the residual quantization noise, after MA filtering and decimating by N, is white noise which has a variance reduced by a factor of N as compared to the variance of the unprocessed quantization noise.

4.2.1 Effect of correcting passband amplitude on residual noise

When the amplitude of the passband is equalized (increased), which is necessary due to the non-flat passband of the MA filter, the average power of the residual quantization noise is increased. From (24), the final SNR of the entire system is,

$$\mathrm{SNR}_{tot} = \frac{\sigma_x{}^2 \, N}{(\sigma_{e_q}{}^2)\,(\eta_{gain})} \tag{25}$$

where $\sigma_x{}^2$ is the variance of the input signal, and η_{gain} is the noise power gain as a result of the passband equalization process.

When scaling is used for passband equalization, for sinusoidal inputs,

$$\eta_{gain} = \frac{1}{\left| H_{MA}(e^{j\omega_0}) \right|^2} \tag{26}$$

where ω_0 is the frequency of the input sinusoid.

When inverse filtering is used,

$$\eta gain = \frac{N}{2\pi} \int_{\frac{-\pi}{N}}^{\frac{\pi}{N}} \frac{1}{|H_{MA}(e^{j\omega})|^2} d\omega \tag{27}$$

for a given filter length N. Note that the noise after inverse filtering is no longer white.

If the alternative scheme of inverse filtering is used, as described in section 3.2, then

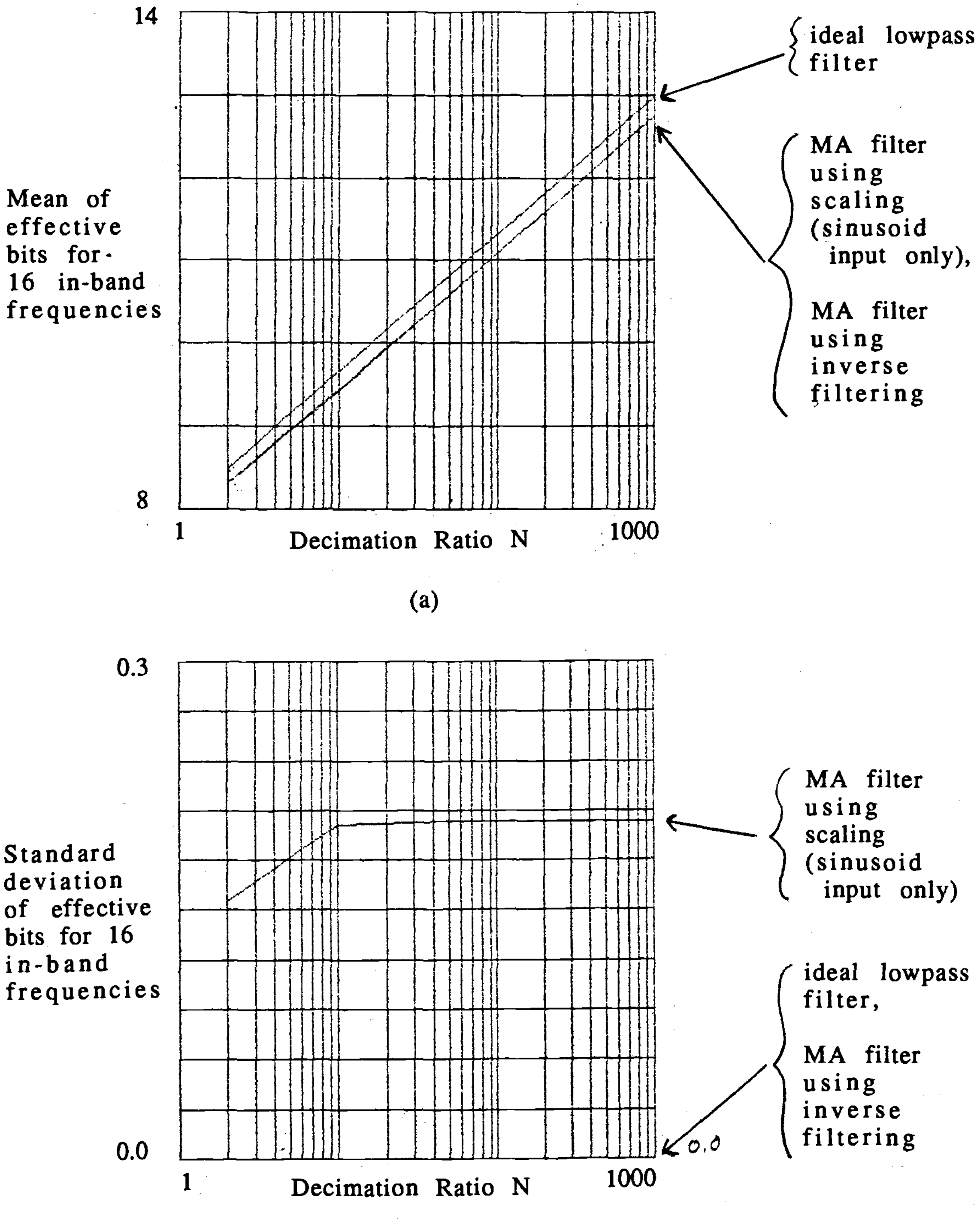
$$\eta gain = \frac{N}{2\pi} \int_{\frac{-\pi}{N^\beta}}^{\frac{\pi}{N^\beta}} \frac{1}{|H_{MA}(e^{j\omega})|^2} d\omega \tag{28}$$

where β is the fraction of the MA passband covered by the signal.

The effective bits can be calculated by,

$$EB = B - \frac{1}{2} \log 2\left(\frac{\eta gain}{N}\right) \tag{29}$$

When scaling is used, $\eta gain$ and EB are dependent on the input signal frequency and EB will decrease with increasing frequency. Since there can be a variation of EB for a given N, defining some average EB for the input signal frequencies over the passband is of interest. To be consistent with the simulation results, the sample mean of EB for 16 input sinusoids equally spaced in frequency over the passband is used as this average measure. Note that using these calculations there is no variation of $\eta gain$, thus EB, with input frequency when inverse filtering is used. A graph of the sample mean of effective bits as a function of N using these theoretical results is shown in figure 20. The sample standard deviation of EB for the 16 input sinusoids, as a function of N, is also shown. Figure 20 also includes



(a)

(b)

Figure 20

Theoretical results assuming white quantization noise, where the raw quantizer is 8-bits. (a) Average resolution enhancement. (b) Standard deviation of effective bits over 16 input frequencies equally spaced over the passband.

theoretical results for the case of using an ideal lowpass filter ($\eta gain = 1$) instead of the MA filter.

4.2.2 Effect of adding white system noise

When white system noise, w(n) in figure 10b, is added to the input signal the effective bits can be calculated by,

$$EB = B - \frac{1}{2} \log 2\left(\frac{\eta gain}{N}\left(\frac{\sigma_{e_q}^2 + \sigma_w^2}{\sigma_{e_q}^2}\right)\right) \tag{30}$$

and the SNR by,

$$SNR_{tot} = \frac{\sigma_x^2 N}{(\sigma_{e_q}^2 + \sigma_w^2)(\eta gain)} \tag{31}$$

where σ_w^2 is the variance of the additional system noise. Here it has been assumed that the quantization noise is uncorrelated with the additional system noise. The theoretical results of (30), assuming an ideal lowpass filter, are graphed, along with simulation results in figures 25-30 for cases of different noise variances.

4.3 Comparison to noise shaping systems which use the MA filter

As seen in figure 6, noise shaping coders offer much better resolution enhancement over the simpler lowpass method when an ideal (or near ideal) lowpass decimation filter is used. However, to achieve high operational speed a simple filter is required. Figure 21b shows the average resolution enhancement for various noise-shaped cases and the non noise-shaped case where the MA filter is used as the lowpass filter. The white noise model is assumed. In this figure the transfer function from the noise source to the input of the decimation filter is

$$T(z) = \left(1-z^{-1}\right)^{n} \tag{32}$$

for nth order noise shaping. Figure 21a shows the magnitude of T(z) for n of 1, 2, and 3. Figure 21b clearly shows that using simple filters can drastically reduce the performance of noise shaping coders. Therefore, considering a cost/performance analysis, lowpass filtering could be a better alternative to noise shaping coders at high input sample rates.

4.4 Worst case performance

An input signal for which the proposed system has the poorest resolution enhancement is a constant-valued signal (other signals could also cause the same performance, but never poorer performance). In this case the quantization error is also constant-valued. Therefore, lowpass filtering will have no effect on the quantization error and there will be no resolution enhancement. Note that the resolution is not degraded from the case of using the quantizer alone; this was one of the objectives of the thesis. As discussed in chapter 2, dither can be used to significantly improve the resolution for input signals which otherwise cause poor performance.

This chapter can be concluded by two basic statements. First, from figure 20, we expect the lowpass system using the MA filter to perform within $\frac{1}{4}$ of an effective bit of an ideal lowpass system. Second, from figure 21b, when the MA filter is used as the lowpass decimation filter (applicable to high sample rates), the simpler lowpass system provides at least $\frac{1}{2}$ the resolution enhancement (in effective bits) as a noise shaped coder.

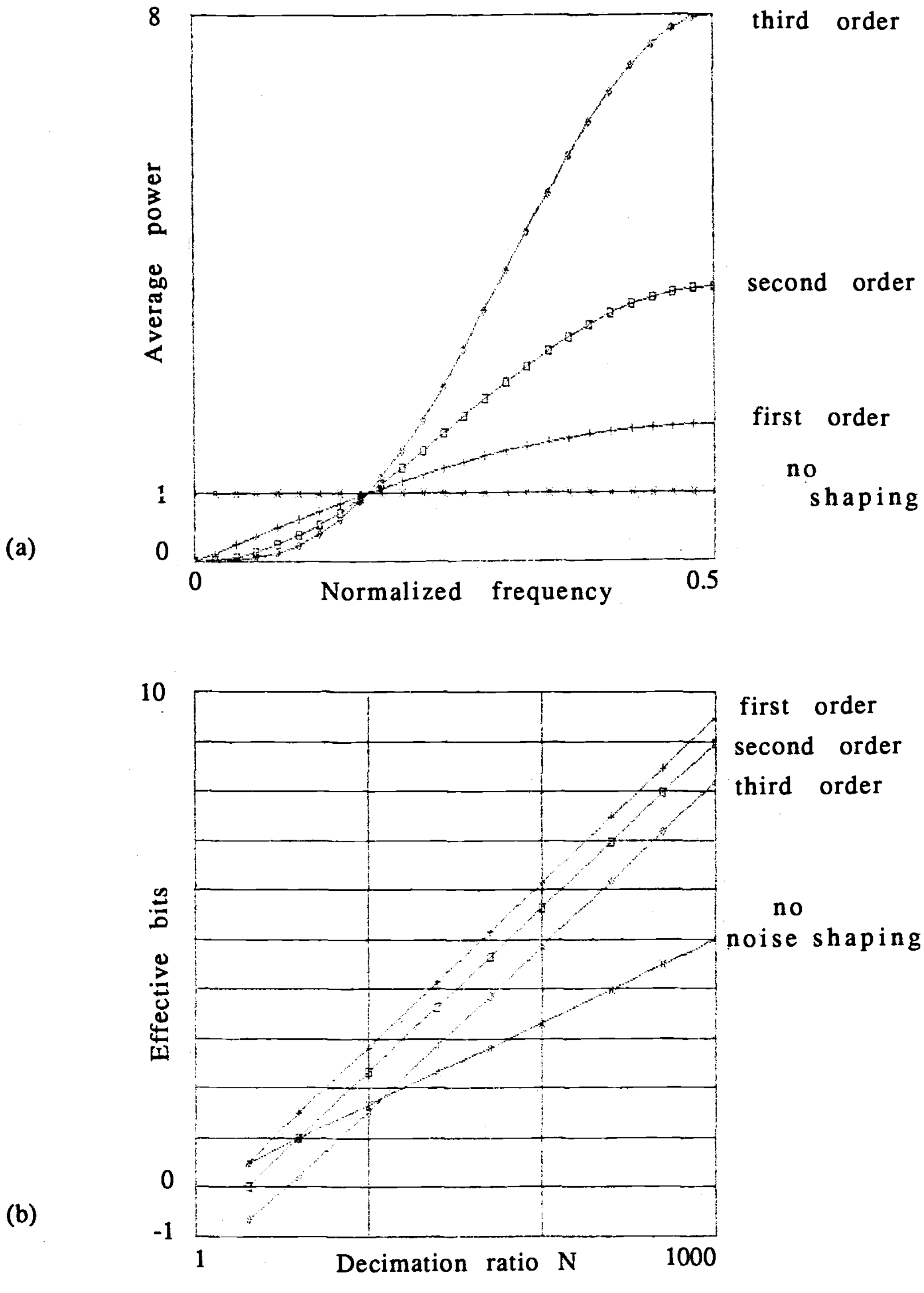


Figure 21

(a) Effect of noise shaping on white noise.
(b) Resolution enhancement for different order noise shaping coders where MA filter is used as final lowpass filter--assuming white quantization noise.

Chapter 5

Simulations

In this section, the results from an extensive set of simulations is presented which demonstrates the resolution enhancement of the MA filter lowpass system. Simulations were performed using sinusoidal inputs with and without the addition of white system noise. These simulation results are used to measure effective bits while avoiding alias errors. Simulations were also performed using non-bandlimited pulse train inputs. Simulations were performed with and without the addition of white system noise. These simulation results are aimed at investigating alias error and the beneficial effect of system noise (dither) on square-like waves. A summary of the definitions of error signals, and the methods used to calculate them is presented in Appendix 1.

5.1 Simulations for effective bits measurement

Simulations were performed where the inputs to the model in figure 10b are sinusoidal, ie. $x(n) = A\cdot\cos(\omega_0 n)$, for two basic cases. In the first case, there is no system noise ($w(n) = 0$). In the second case, system noise is added. The signal quantizer was an ideal, uniform, 8-bit quantizer and the output of the decimator, $y_{MA}(nN)$, was normalized and rounded to 16 bits. Both methods of amplitude correction, scaling and inverse filtering, were used. The filtering transients were removed from the output signal before analysis.

The sinusoid frequencies are in the passband of the decimator filter and the signal alias effects after down sampling are not considered. For each moving average filter length (N), 16 sinusoids were generated with normalized frequencies

$$\omega 0_i = \frac{2\pi i}{(32N + \text{fraction})} \quad , i = 1,\ldots,16. \qquad (33)$$

Note that these 16 sinusoids divide the decimation filter's passband ($|\omega C| < \frac{\pi}{N}$) into 16 nearly equal parts. The "fraction" term in (33) was included so that the period of the sampled sinusoid would be increased. If this term were not included, the quantization error would contain large magnitudes at harmonics of the input sinusoid's frequency, in which case the white noise model for the quantization error would be poor.[6]

The output record length was 512 samples after all transients from MA filtering and inverse filtering were removed and decimation was performed. These sinusoids all had a magnitude such that they covered 0.95 of the full scale range of the quantizer.

Two methods of calculating effective bits (EB) were used.

Tektronix Effective Bits Package[7] : In this method, the input to the system must be a sinusoid. The output sequence from the system ($A[y_{MA}'(nN)]$ in figure 10b) is fed directly to the program. The frequency of the input sinusoid is also reported to the program. This program fits the sequence (least squares) to a sinusoid of the same frequency. The error sequence is calculated as the differences between the system output data sequence and the least-squares-fit sinusoidal sequence. Effective bits is then computed by substituting the error average power, into the U_q' variable in the effective bits formula.[8]

[6] For simulation case where "fraction" = 0, see Appendix 3.

[7] This package was made available by Tektronix. It was written by Marc Frajola and Dan Knierim--both of Tektronix. It uses the method described in [13].

[8] Equation (A31) in Appendix 2.

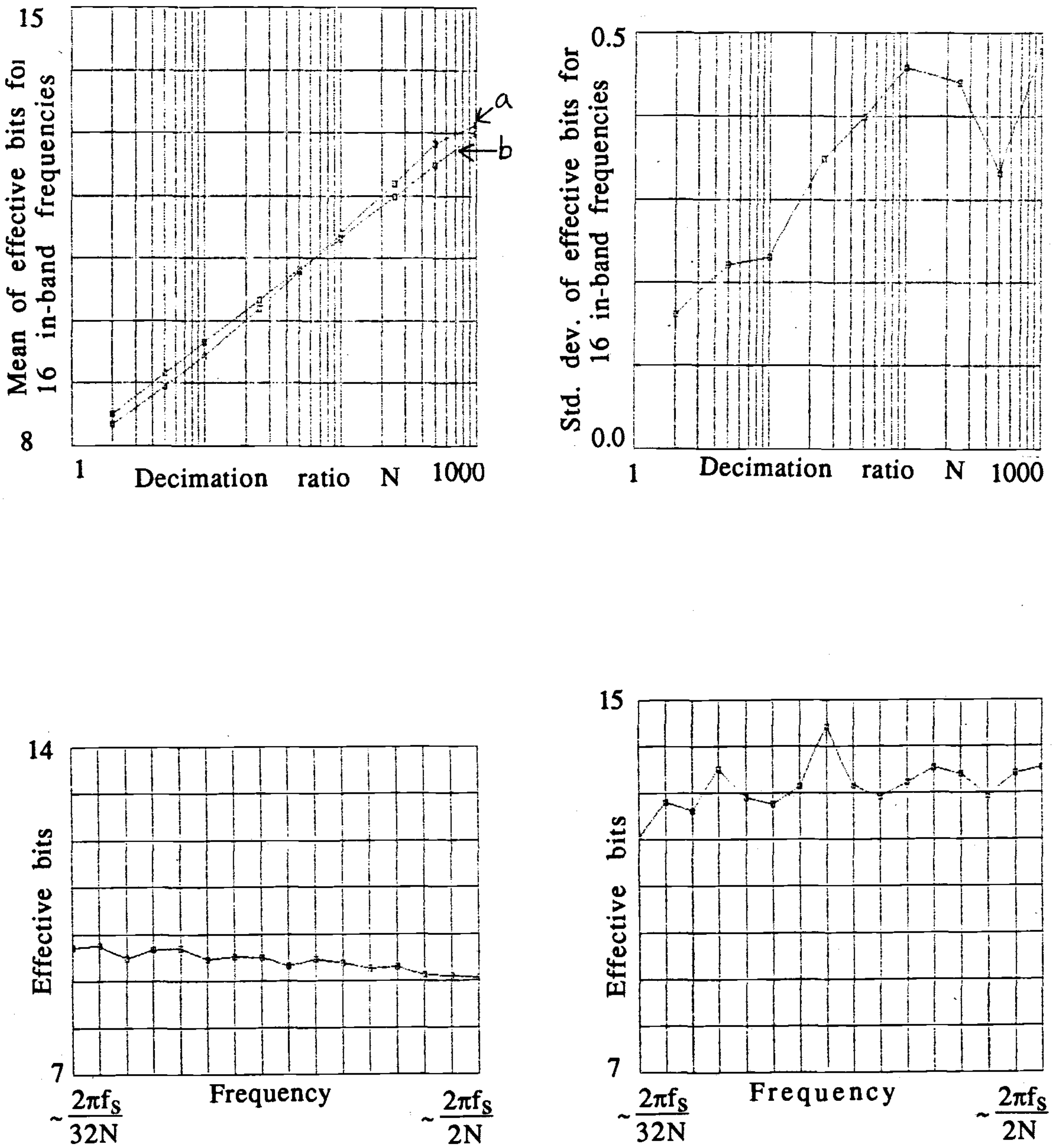
Direct calculation of error average power: The second method is described in detail in Appendix A1.4. This method calculates the final error signal by feeding the actual quantization error sequence[9] $e_q(n)$ through the system shown in figure 10b. The error average power is substituted into the U_q' variable for the effective bits formula.[10]

5.1.1 Simulation results for the case of no system noise

Results for the case of using the Effective Bits Package and performing amplitude correction by scaling are summarized in figure 22, where $w(n) = 0$. The most important graph in figure 22 to consider is the graph of the mean effective bits as a function of N. Comparing the simulation results in figure 22 to the theoretical results in figure 20 it can be seen that for small N the simulated results of mean EB are nearly the same as the results using the theoretical model. However, as N increases beyond 10 the mean EB starts to increase at a rate greater than anticipated by the theoretical results. At N > 70, the simulated EB actually becomes greater than the theoretical results using an ideal low pass filter. The two graphs at the bottom of figure 22 show the variation of effective bits with input signal frequency. In the graph where N=1000, note that for the most highly oversampled signals (low frequency part of band) the effective bits decreases. These phenomena can be attributed to the breakdown of the assumption that the quantization error is uncorrelated (white noise). The model for quantization noise for deterministic inputs presented in chapter 2 predicts that the noise will become non-white for higher N. From (8) we have $N \ll \pi\alpha A$ for the noise to be white, where $0 \leq \alpha \leq 1$ represents the fraction of the MA filter passband of the input sinusoid frequency $(\alpha = \frac{\omega_0 N}{\pi})$. The lowest frequency sinusoid (of the 16 used) for a given N corresponds to $\alpha = \frac{1}{16}$. For

[9] $e_q(n) + w(n)$ if system noise is non-zero. $x(n)$ is held at zero.

[10] Equation (A31) in Appendix 2.



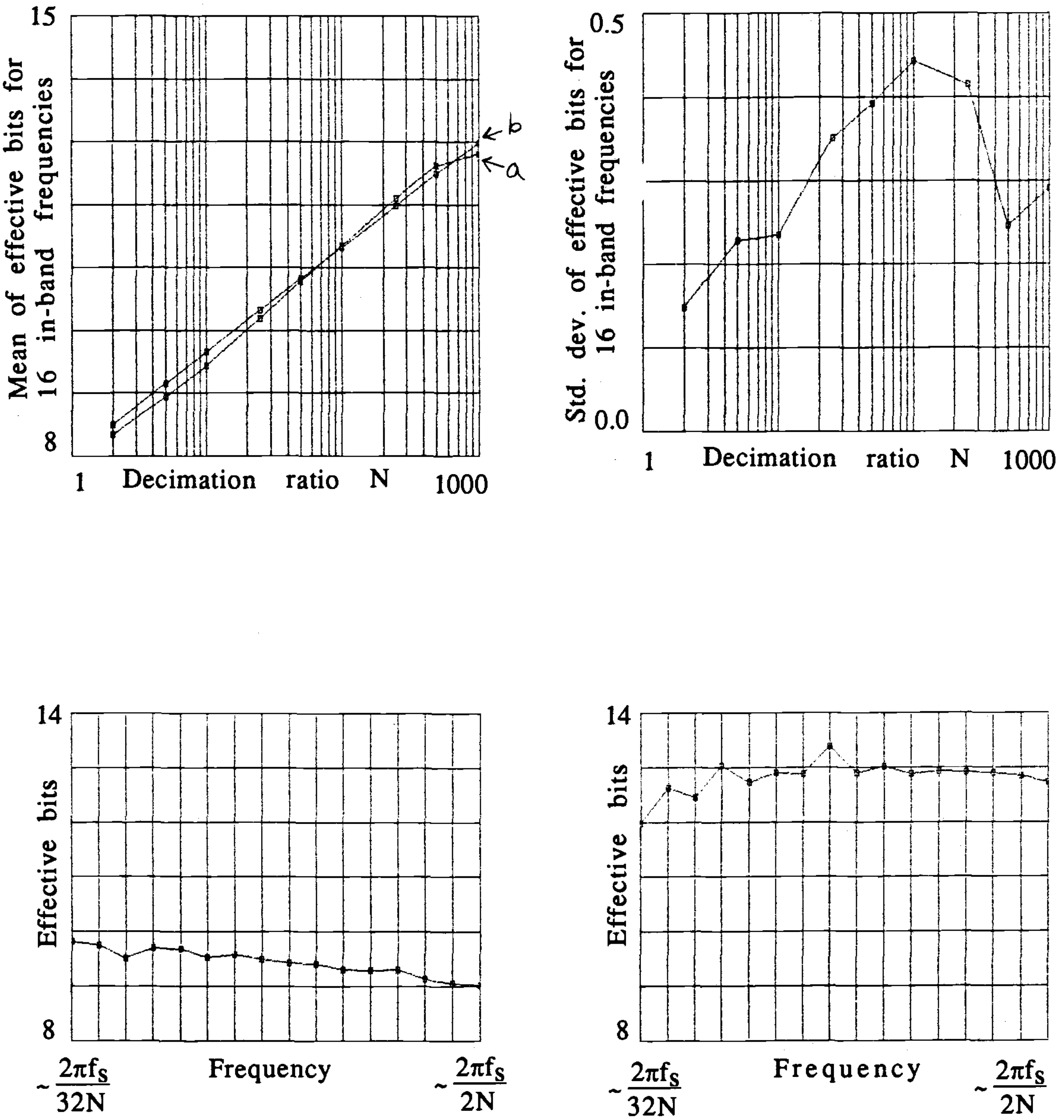
N = 10

N = 1000

a - Simulated results using MA decimation filter.
b - Theoretical results using an ideal lowpass filter and white quantization noise models.

Figure 22

Effective bits for sinusoidal inputs with no noise. Passband amplitude corrected by scaling. Effective bits computed using Tektronix Effective Bits Package.



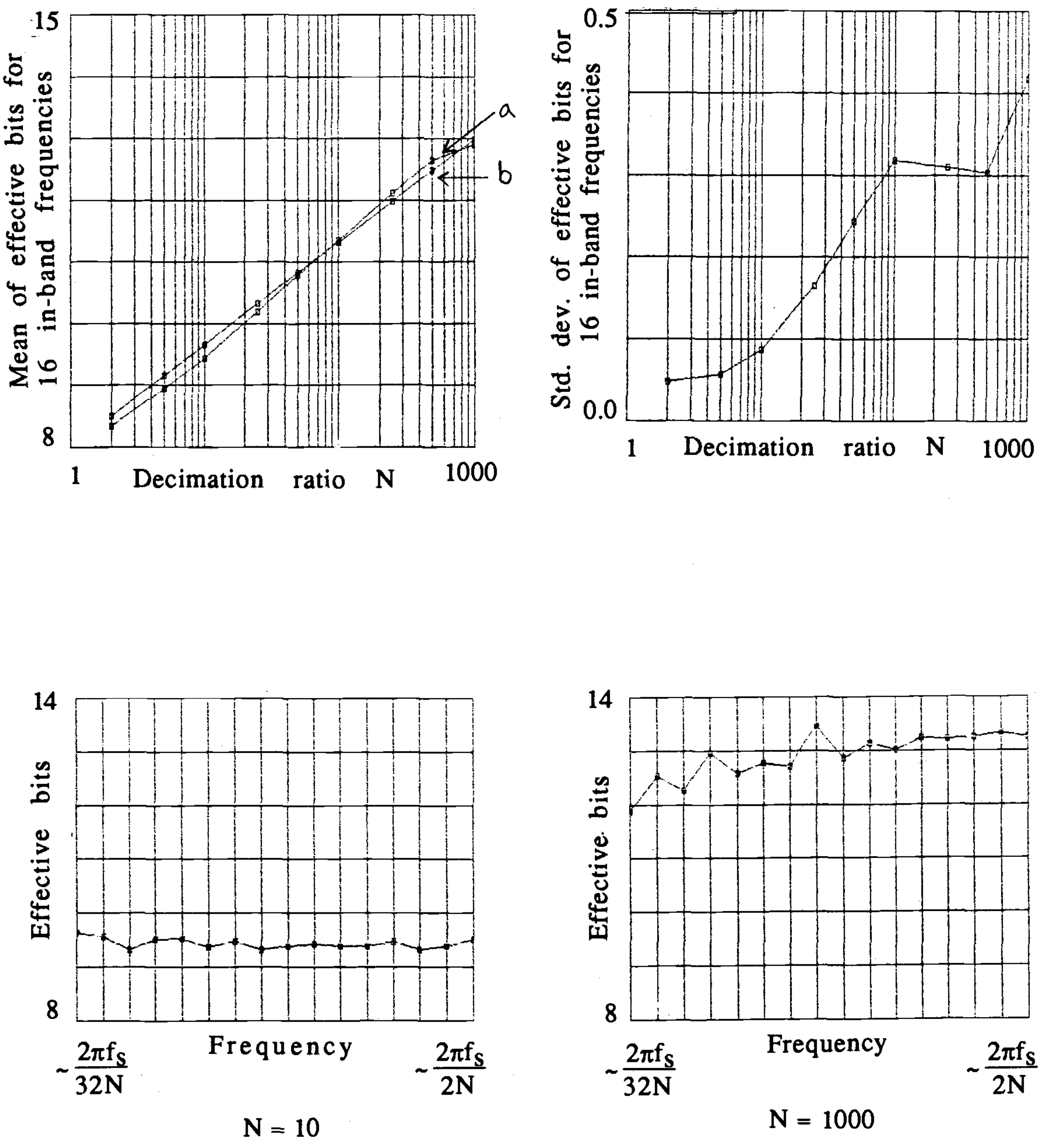
N = 10

N = 1000

a - Simulated results using MA decimation filter.
b - Theoretical results using an ideal lowpass filter and white quantization noise models.

Figure 23

Effective bits for sinusoidal inputs with no noise. Passband amplitude corrected by scaling. Effective bits computed using quantization error sequence directly.



a - Simulated results using MA decimation filter.
b - Theoretical results using an ideal lowpass filter and white quantization noise models.

Figure 24

Effective bits for sinusoidal inputs with no noise. Passband amplitude corrected by inverse filtering. Effective bits computed using quantization error sequence directly.

these simulations, A=128·0.95. Therefore the model predicts that for N<<24 the noise should will be white. As N approaches 24, the lowest frequency sinusoid would produce non-white quantization noise. As N approaches 384, all 16 sinusoids will produce non-white noise.

Comparing the graphs of the standard deviation of EB vs. N in figure 22 and the corresponding graph of the theoretical results in figure 20 it is clear that the simulation results do not correspond well for large N. Simulation results agree better when system noise is added, as will be seen later.

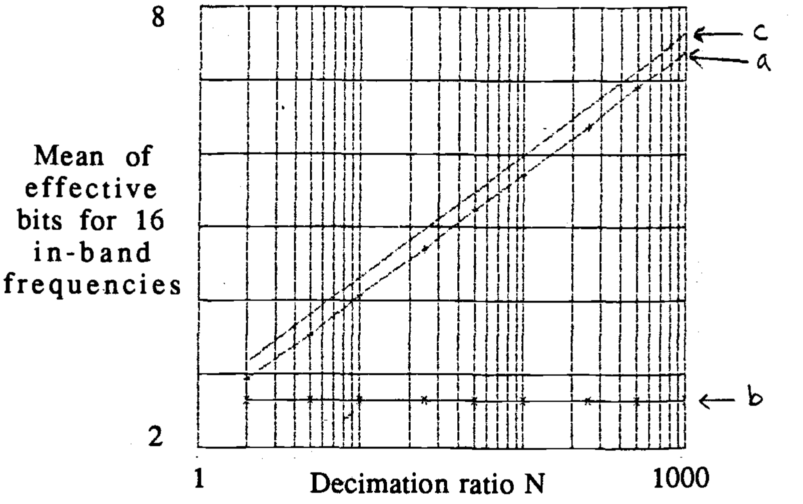
When the quantization error sequence is used directly to calculate EB, slightly different results are obtained. These results are summarized in figure 23. The main difference between the results of the two methods is that for N > 100 the mean and the variance of the EB over the passband are slightly less for the method using the error sequence. The fact that the mean EB is higher when using the Effective Bits Package makes sense since, by definition, the least squares sinusoid of all possible sinusoids of the known frequency is found from the quantized samples. This is used as if it were the actual input sinusoid. Therefore, "least squares" translates to "highest EB" of all sinusoids of the known frequency.

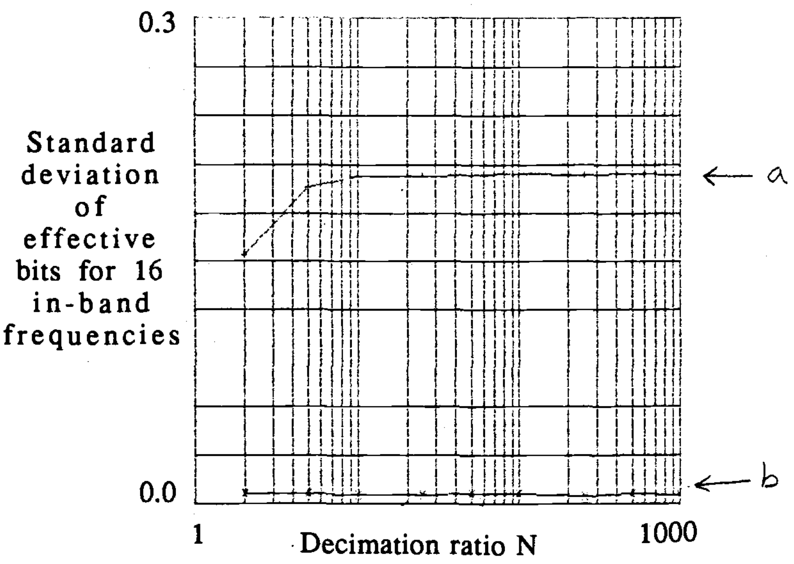
Results obtained when inverse filtering is used for amplitude correction are summarized in figure 24. Recall that the theoretical model predicts that there should be no deviation of EB across the passband when this method is used. In the simulations there is significant deviation. As will be seen later, when system noise is added, this deviation is reduced.

5.1.2 Simulation results for different cases of system noise

In all cases the noise w(n) was a uniformly distributed pseudo random sequence. Three cases of noise amplitude ranges were simulated,

$$\frac{A_w}{A_x} = \{0.2, 0.02, 0.002\} \qquad (34)$$

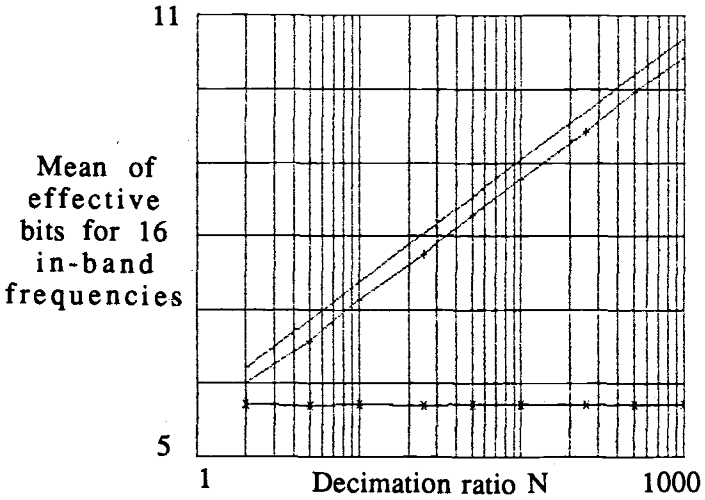


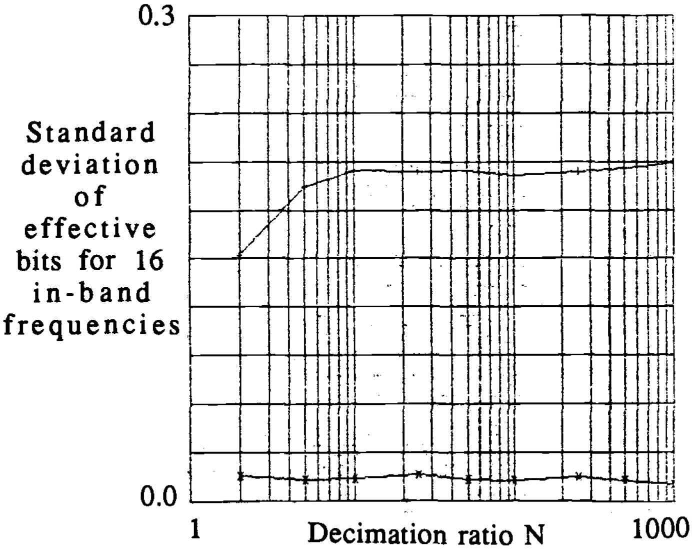


a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 25

Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.2 · sine amplitude range). Scaling is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.

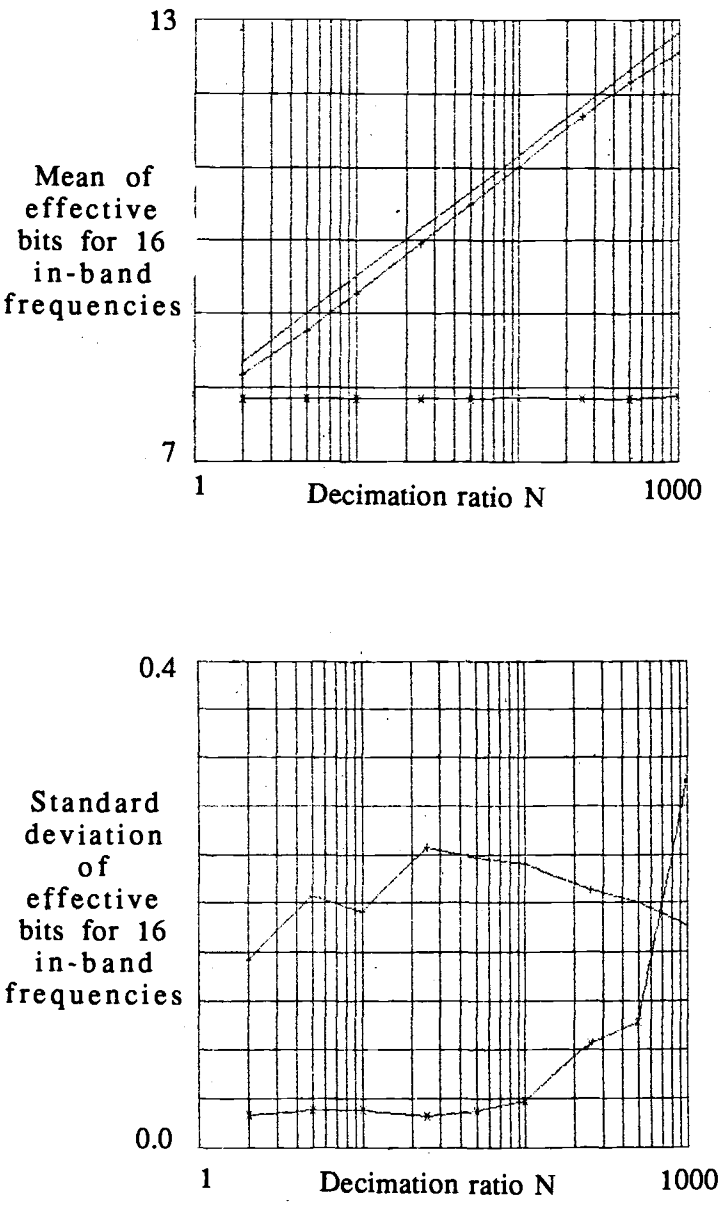




a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 26

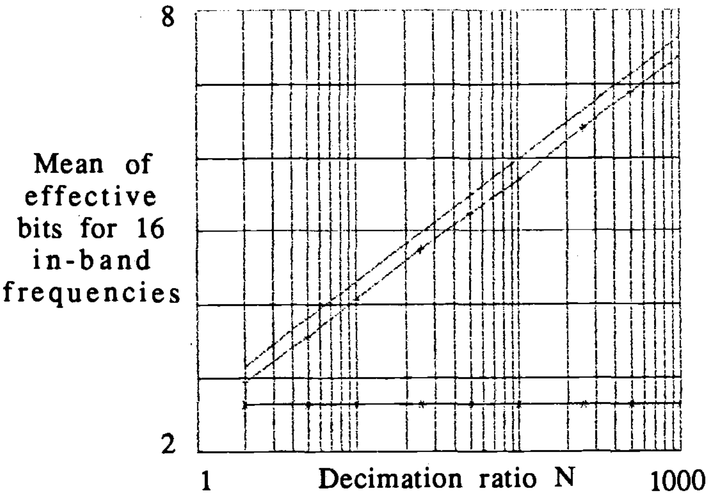
Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.02 · sine amplitude range). Scaling is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.

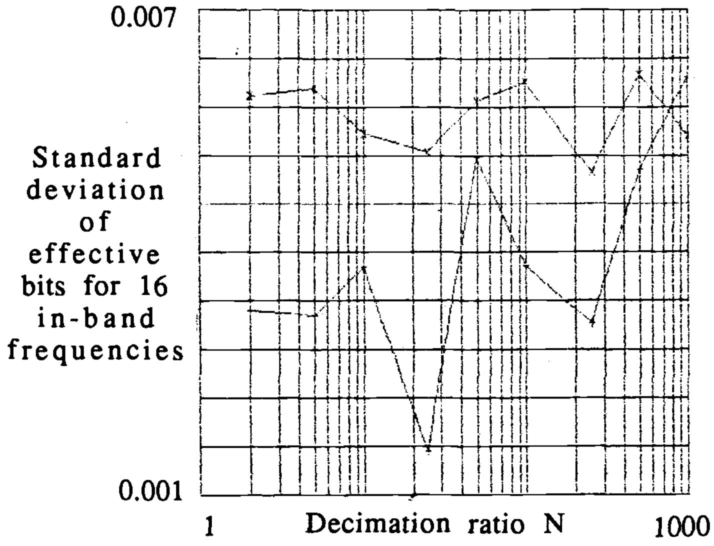


a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 27

Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.002 · sine amplitude range). Scaling is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.

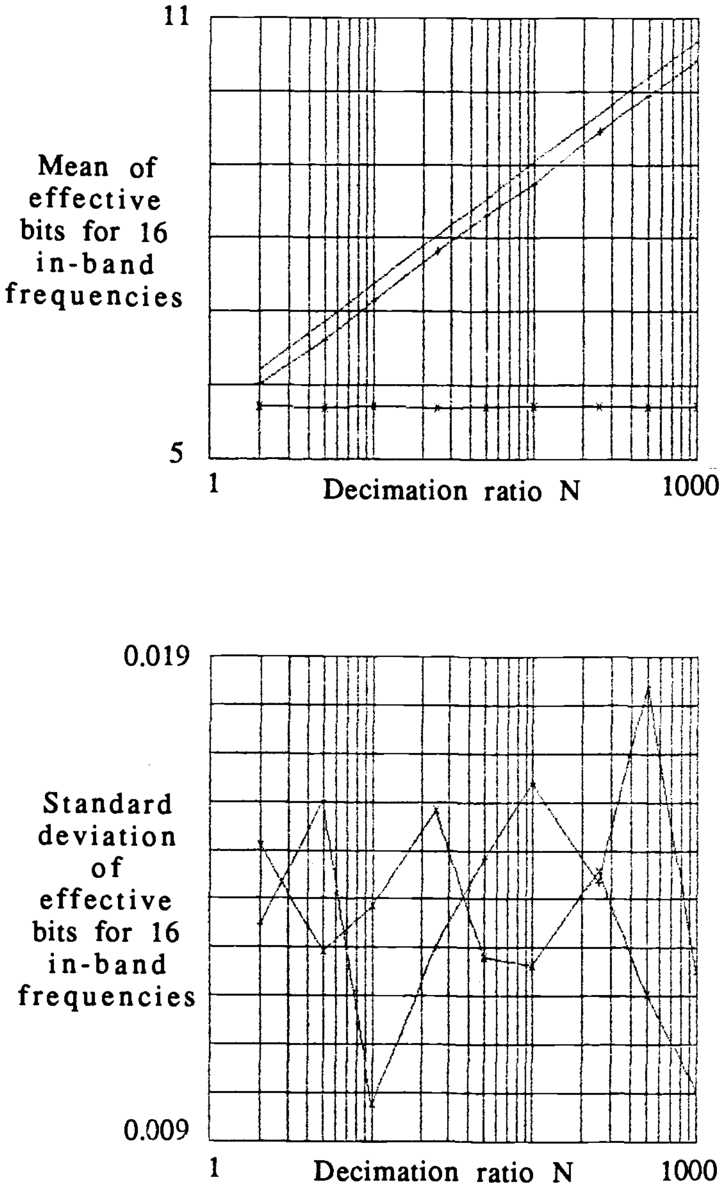




a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 28

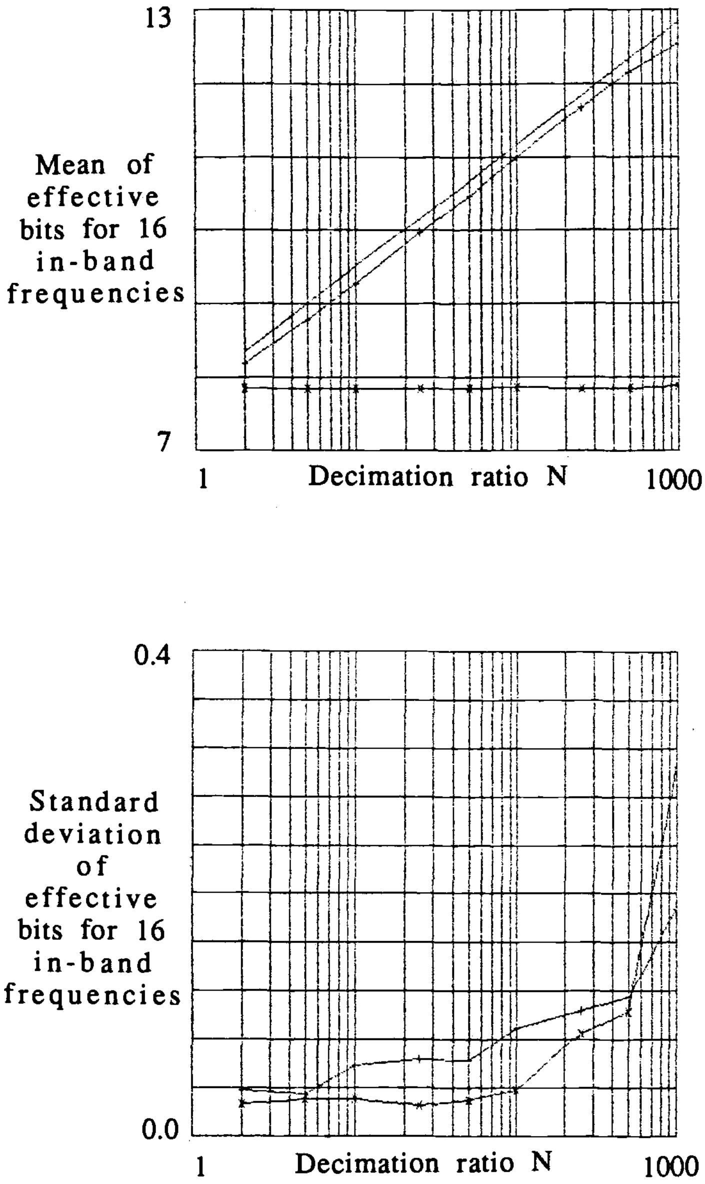
Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.2 · sine amplitude range). Inverse filtering is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.



a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 29

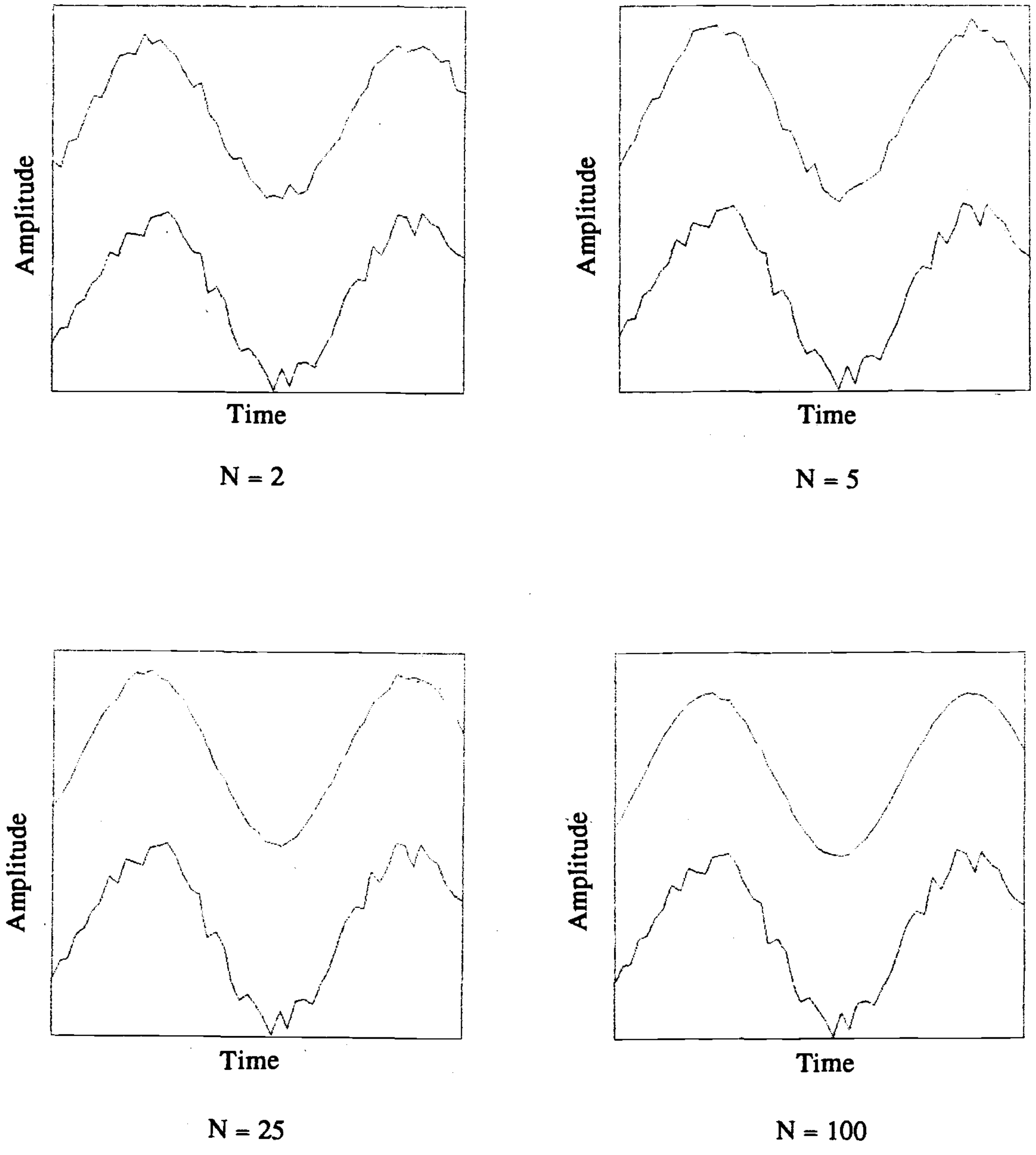
Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.02 · sine amplitude range). Inverse filtering is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.



a - Simulated results using MA decimation filter.
b - Simulated results without filtering.
c - Theoretical results using an ideal lowpass filter and white noise model for quantization and system noise.

Figure 30

Effective bits for sinusoidal inputs, with uniform random system noise added of amplitude range (0.002 · sine amplitude range). Inverse filtering is used to correct MA filter passband amplitude. The raw quantizer is 8-bits.



N = 2

N = 5

N = 25

N = 100

Figure 31

Time domain graphs of a noisy sinusoid before and after MA filtering. In each window the bottom graph is before filtering and the top graph is after filtering. The noise amplitude range is (0.2 · sine amplitude range).

Signal frequency is ~$2\pi f_S/32N$. Output is decimated by N. MA filter amplitude roll-off corrected by scaling.

where A_W is the amplitude range of w(n), and A_X is the amplitude range of x(n), the sinusoidal input. These noise amplitudes correspond to {± 20 LSB, ± 2.4 LSB, ± 0.24 LSB} of the 8-bit quantizer, respectively. The SNR can be determined by,

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_w^2} = \frac{\left(\frac{\left(\frac{A_x}{2}\right)^2}{2}\right)}{\left(\frac{A_w^2}{12}\right)} = \frac{3}{2}\frac{A_x^2}{A_w^2} \qquad (35)$$

where σ_x^2 is the average power in the sinusoidal input x(n), and σ_w^2 is the average power in the system noise w(n). Therefore, the cases of noise amplitude in (34) correspond to SNRs of {15.7 dB, 35.7 dB, and 55.7 dB}, respectively.

Both methods of amplitude correction were simulated. The results where scaling was used are summarized in figures 25-27. The results where inverse filtering was used are summarized in figures 28-30. Effective bits was computed using the Tektronix Effective Bits Package for all cases.

The mean EB over the band are essentially identical for both amplitude correction methods. However, the deviation of EB over the band is less when inverse filtering is used.

Comparing the white noise model results for (mean EB) vs. N in figure 20 to the simulation results where w(n) = 0, figures 22-24, and to simulation results where system noise is added, figures 25-30, it is seen that the addition of system noise causes the simulated results to agree more closely with the theoretical white noise results. This is consistent with the theoretical results of section 2.3 which show that dither signals tend to whiten the quantization noise.

When system noise is present, the mean EB is generally about 1/4 of an effective bit less than the theoretical results for the case of ideal lowpass filtering, regardless of N. Also note that the deviation of the EB over frequency, for a given N, is reduced when noise is added.

The ability of the MA filter to remove noise from a sinusoid for the case of 15.7 dB SNR, seen in the time domain, is demonstrated in figure 31.

5.2 System performance in reducing alias error for non-bandlimited signals

In previous sections, the input signal x(n) was a sinusoid and was always bandlimited within the passband of the decimation filter.[11] Now the case is considered where the input signal is not constrained to the frequency range of the filter passband. First, the undesirable effects of aliasing and high frequency signal cutoff will be discussed. Next, the trade-offs between reducing aliasing error, residual quantization error, and passband attenuation error, depending on the configuration of the system are described. Finally, simulation results are presented for example inputs from a class of non-bandlimited signals. The last simulation results presented examine the beneficial effects of system noise (dither) for certain signals.

5.2.1 Alias and cutoff error for non-bandlimited inputs

When performing decimation, aliasing is avoided if the signal components having frequencies greater than the folding frequency[12] are eliminated.

[11] The system noise added to the input, w(n), extended beyond the cutoff frequency of the decimation filter. But we are not concerned with aliasing effects on the noise--only its average power.

[12] The folding frequency is, $\omega_c = \frac{\pi f_s}{N}$, where N is the decimation ratio.

The decimation filter is used to approximate this result. However, in this case, the high frequency components of the signal are filtered out, this could also be considered an error. Therefore two separate sources of error result when the non-bandlimited signal is sampled, filtered, and decimated: alias error resulting from non-ideal stopband characteristics of the decimation filter, and cutoff error caused by the fact that the signal components at frequencies greater than the folding frequency are not represented. The alias error can be decreased by using a lowpass filter which is a better approximation to an ideal lowpass filter. The cutoff error is, in general, unavoidable for non-bandlimited inputs--the only way to improve it is increase the sampling rate. The cutoff error is not considered in this study. However, the alias error is. Figure 32 clarifies the difference between these two separate error sources.

5.2.2 System configuration trade-offs for reducing error

Three system configurations will be considered. Simulations were performed for all three cases which will be discussed in the next subsection.

In the first case there is no signal processing. The input signal is quantized, then decimated. For this configuration there is no passband attenuation error, but alias error could result from the decimation process. There is no reduction of the quantization error. The second case uses $y_{MA}'(nN)$ from figure 10b as the system output. That is, no amplitude correction is performed. In this case the passband attenuation error must be considered. The MA filter will significantly reduce the alias error and the quantization error. The last case uses $A[y_{MA}'(nN)]$ from figure 10b as the system output. That is, amplitude correction is performed so that the passband attenuation error is eliminated[13] . However, the effect of

[13] In an actual implementation it would not be completely eliminated. If scaling were used for a sinusoidal input, the exact frequency of the sinusoid would not be known, and the input would not be a perfect sinusoid. If inverse filtering were used, there would always be some difference

correcting the passband attenuation will increase the residual quantization error, $e_q'(nN)$, and the residual alias error, $e_a'(nN)$.

5.2.3 Description of simulations using periodic non-bandlimited inputs

The class of exponentially damped pulse trains were used to represent signals which are not band-limited. By varying the amount of damping we can obtain near-triangle and square waves. Time domain plots of example pulses used in the simulations can be found in figure 33. For each simulation case, 50 different pulse train inputs were generated having fundamental frequencies equally spaced covering 1/10 of the passband of the MA filter. Each input record contained one or more exact periods of the sampled input, thereby validating the method described in appendix A1.3.2. In all cases it was assumed that the input signal, before decimation, was sampled without aliasing. A decimation factor of 10 was used for all simulations. All signals had an amplitude range of -1.0 to 1.0. The input quantizer amplitude range was from -1.05 to 1.05.

5.2.4 Simulation results for the case of no system noise

Damping ratios, τ, of 0.5, 0.2, and 0.05 period were used. The results from these simulations are summarized in figures 34-36. Each figure presents simulation results for pulse train inputs of the same damping ratio. For

between a realizable filter and the ideal inverse filter. In the simulations which follow, inverse filtering was performed by linear phase, 511 tap, FIR filters. However errors were not calculated comparing system input to system output (there would be no way to distinguish different error sources). The main purpose of inverse filtering was to see its effect on increasing the residual quantization noise. Inverse filtering also increases alias error, but this effect was calculated by using the ideal inverse filter response in the frequency domain for a given frequency component of a line-spectra signal.

each damping ratio, the three system configurations discussed in 5.2.2 were simulated. The rms of the various errors are shown versus the fundamental frequency of the pulse train. Effective bits versus fundamental frequency is also included with each simulation set.[14]

[14] Effective bits is defined, here, as the result of summing error signal average powers for all the measured errors, call this U, and using U in place of the U_q variable in the effective bits equation (equation (A31) in Appendix 2). It is assumed that the different errors are uncorrelated.

$U = U_q + U_a$ (no filtering)

$U = U_{q'} + U_{a'} + U_p$ (MA filtering, but no amplitude correction)

$U = U_{q'} + U_{a'}$ (MA filtering, and inverse filter amplitude correction)

See Appendix 1 for definitions of $U_{q'}$, $U_{a'}$, and U_p.

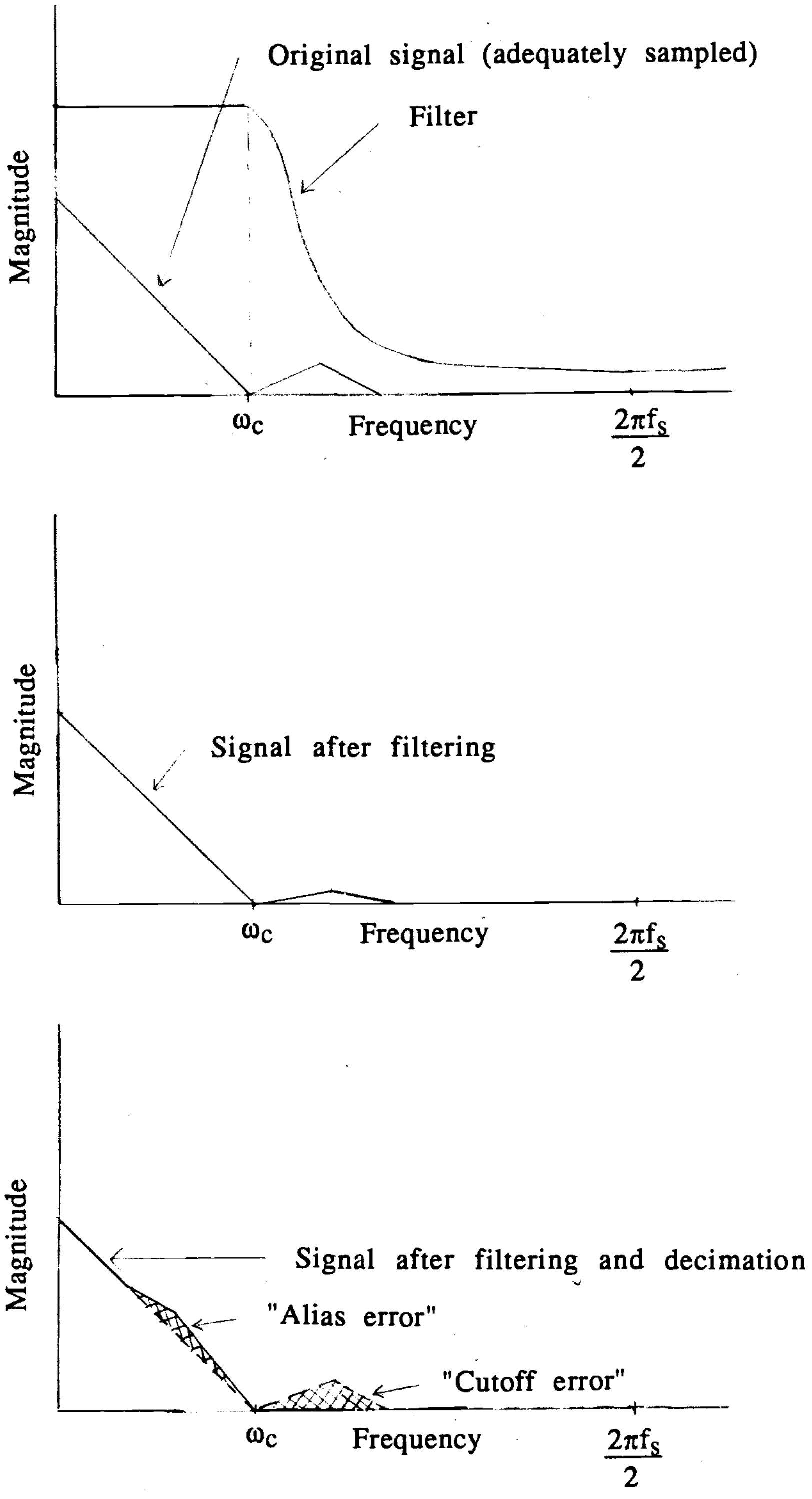
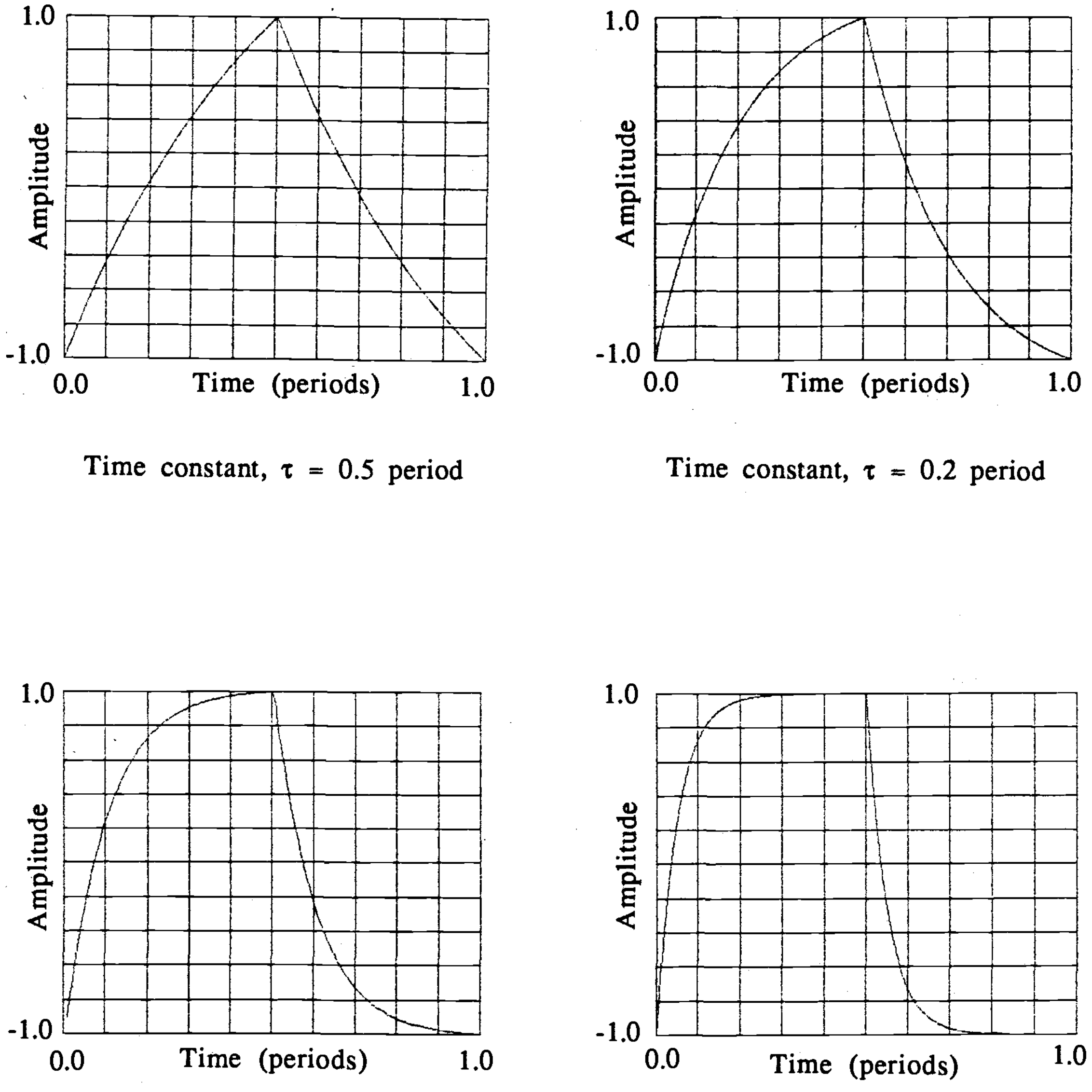


Figure 32

Two errors resulting from the combined operation of filtering and decimation. "Alias error" can only appear at frequencies below $\omega_c = 2\pi f_s/2N$, whereas "cutoff error" can only appear at frequencies above $2\pi f_s/2N$, where N is the decimation ratio.



Time constant, τ = 0.5 period

Time constant, τ = 0.2 period

Time constant, τ = 0.1 period

Time constant, τ = 0.05 period

Figure 33

Time domain graphs of example pulse shapes uses as non-bandlimited inputs in simulations.

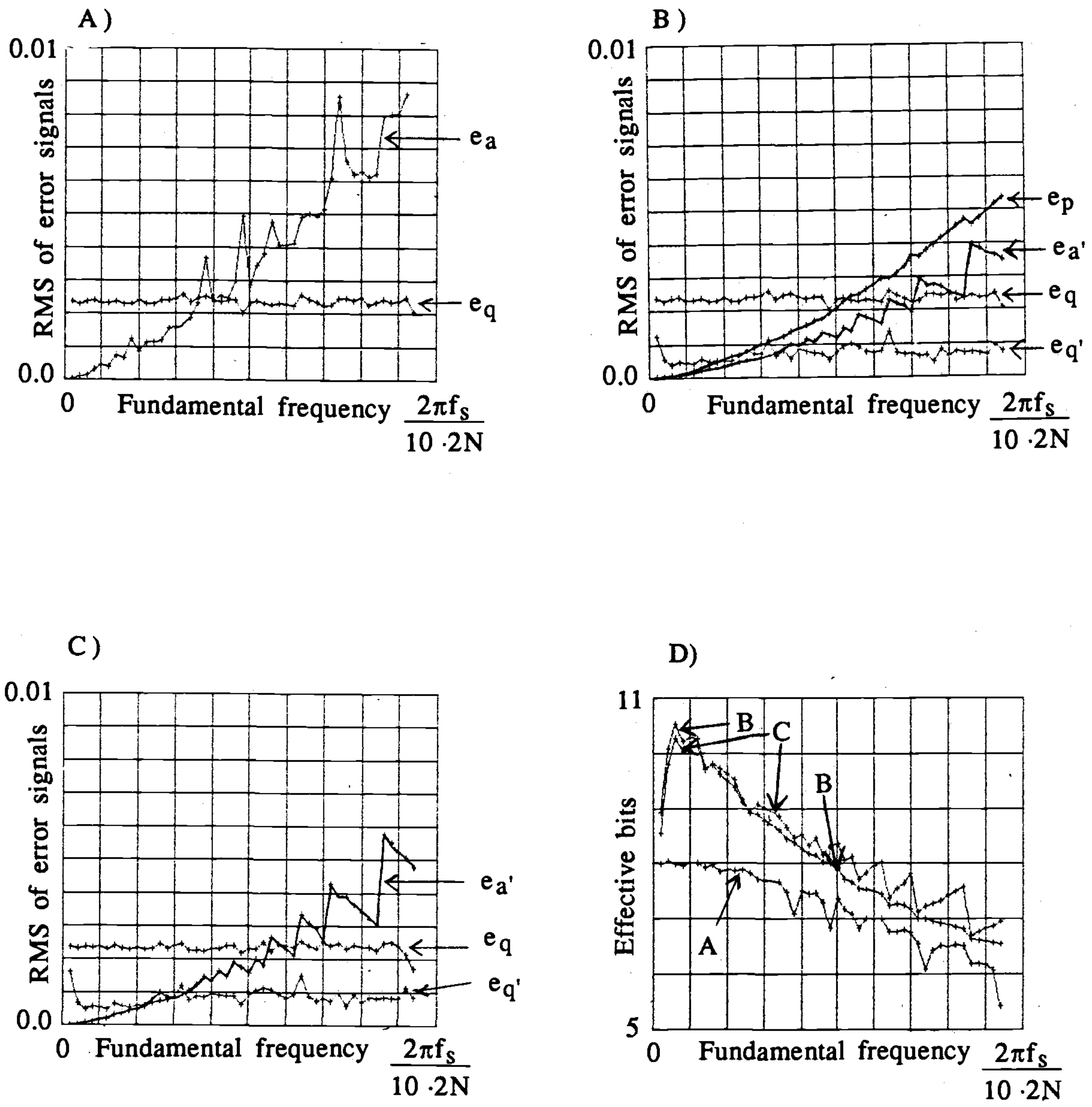


Figure 34

Simulated results for pulse train inputs having a pulse shape of $\tau = 0.5$ period. No system noise is added. Decimation ration N = 10. A/D full scale = 2.1. Pulse amplitude range = 2.0.

A) RMS of errors for case of no filtering.
B) RMS of errors for MA filtering without passband amplitude correction.
C) RMS of errors for case of MA filtering, and inverse filtering.
D) Comparison of effective bits for methods in A), B), and C).

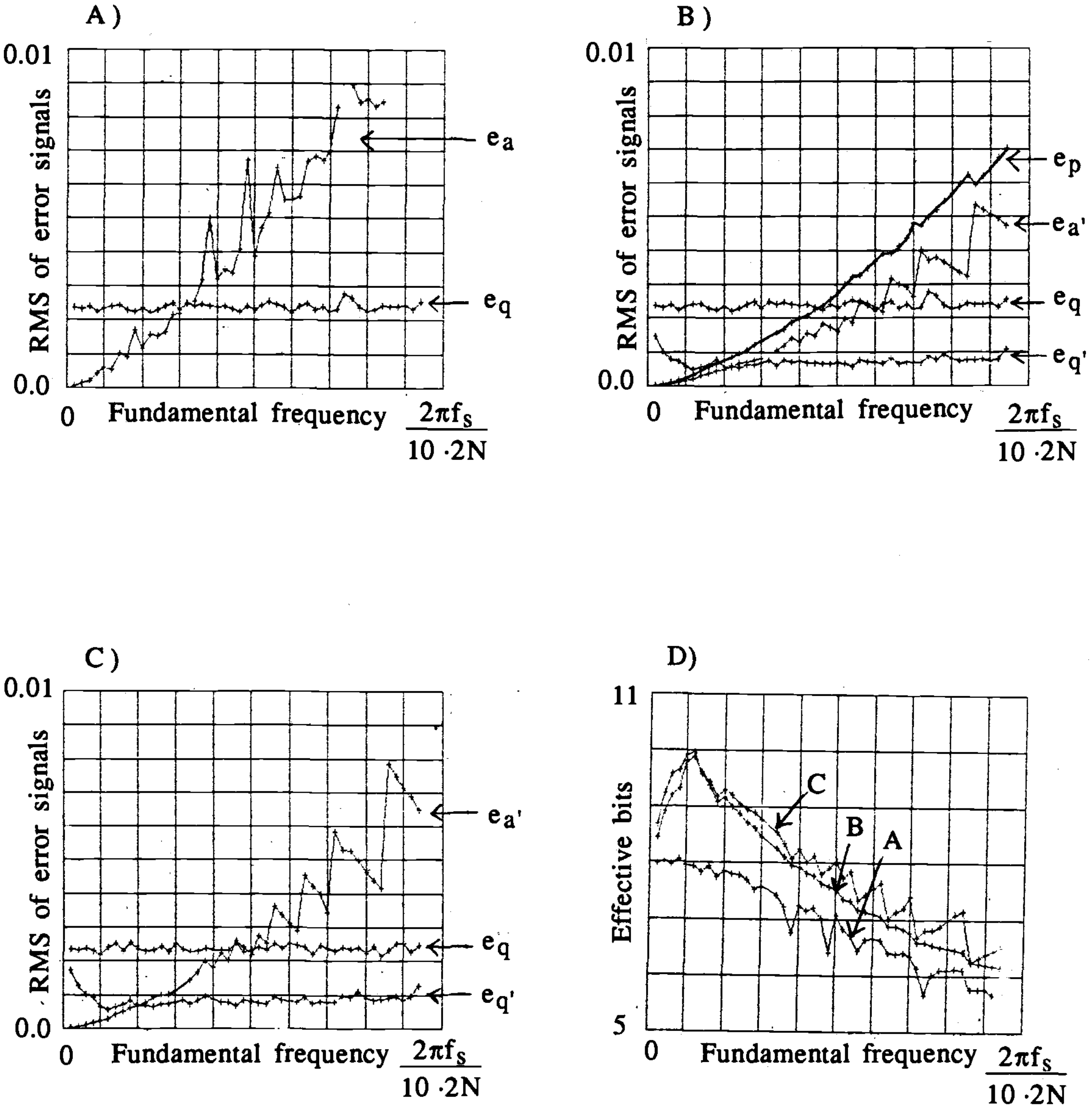


Figure 35

Simulated results for pulse train inputs having a pulse shape of $\tau = 0.2$ period. No system noise is added. Decimation ration $N = 10$. A/D full scale = 2.1. Pulse amplitude range = 2.0.

A) RMS of errors for case of no filtering.
B) RMS of errors for MA filtering without passband amplitude correction.
C) RMS of errors for case of MA filtering, and inverse filtering.
D) Comparison of effective bits for methods in A), B), and C).

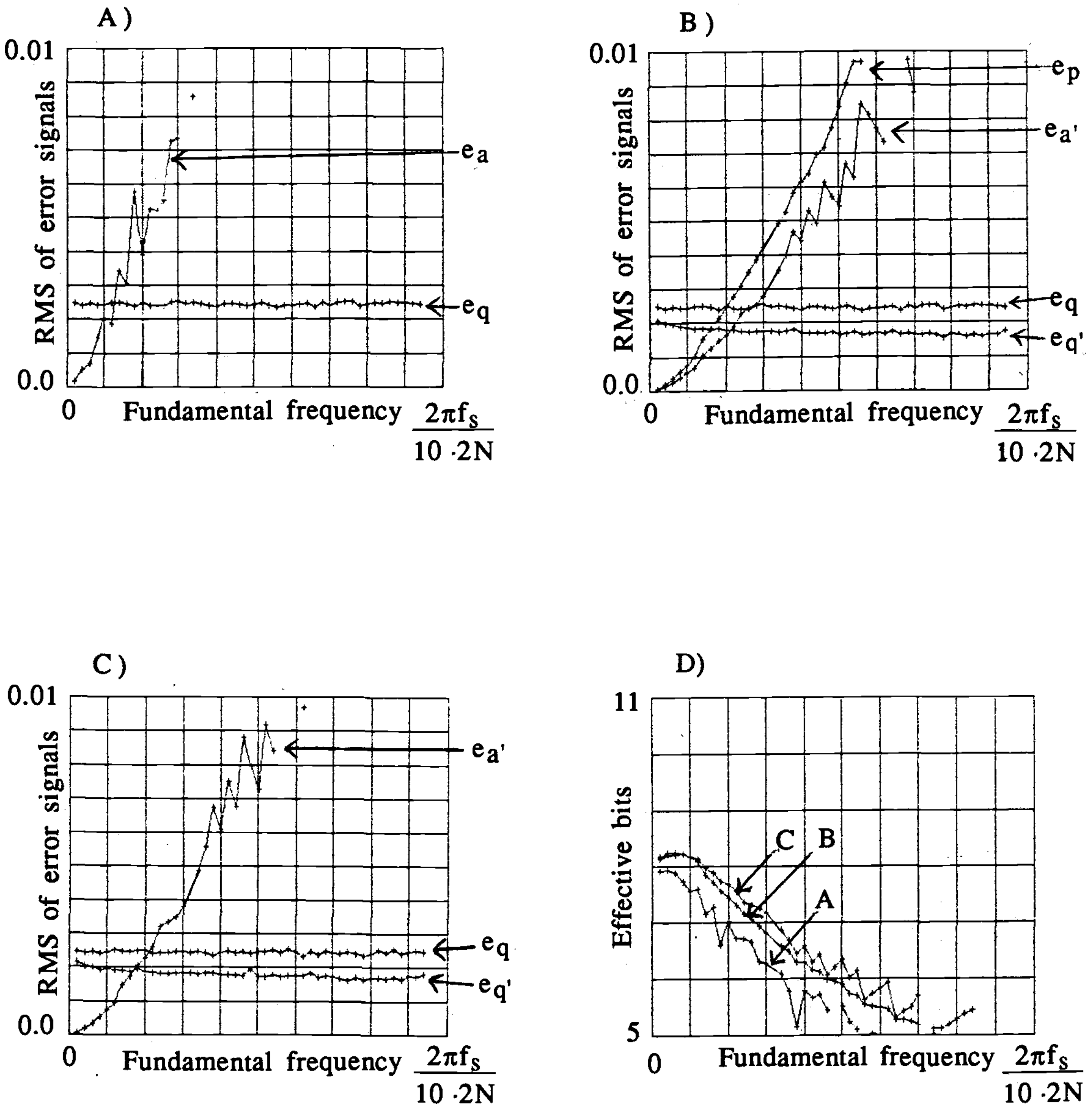


Figure 36

Simulated results for pulse train inputs having a pulse shape of τ = 0.05 period. No system noise is added. Decimation ration N = 10. A/D full scale = 2.1. Pulse amplitude range = 2.0.

A) RMS of errors for case of no filtering.
B) RMS of errors for MA filtering without passband amplitude correction.
C) RMS of errors for case of MA filtering, and inverse filtering.
D) Comparison of effective bits for methods in A), B), and C).

The simulation results for τ = 0.5 period are presented in figure 34, where e_a, e_a', e_q, e_q', and e_p are the rms values of alias error without processing, alias error after processing, unprocessed quantization error, processed (residual) quantization error, and passband attenuation error, respectively. The worst results of the three system configurations is clearly the case of no processing. Of the two system configurations where processing is performed, inverse filtering actually degrades overall performance for very low frequency input signals, as seen in figure 34(D). The results for τ = 0.2 period, shown in figure 35, are similar.

The results for τ = 0.05 period, where the input is nearly a square wave[15] , are shown in figure 36. Here, e_q', is almost as large as e_q. This phenomenon is discussed in the next subsection. The peak in EB for low input frequencies is not present in figure 36(D), as it is in figures 35(D) and 34(D).

5.2.5 Simulation results when system noise is present

When the input is nearly a square wave it is clear that the system is becoming less effective at reducing the rms of the quantization noise[16] . Except for the short transition periods, the signal spends most of the time moving very slowly. In this case, the quantization errors are highly correlated, and the power spectrum of the noise is no longer white; it is predominantly low frequency. So, the lowpass filter is less effective in reducing the noise power. If small amounts of noise (dither) are added to this signal, the quantization noise becomes significantly less correlated.

This idea was verified by a simulations where uniformly distributed pseudo-random system noise, w(n), was added to the input signal. The

[15] See figure 33 for a time domain graph.

[16] Compare e_q and e_q' in figgure 36(B).

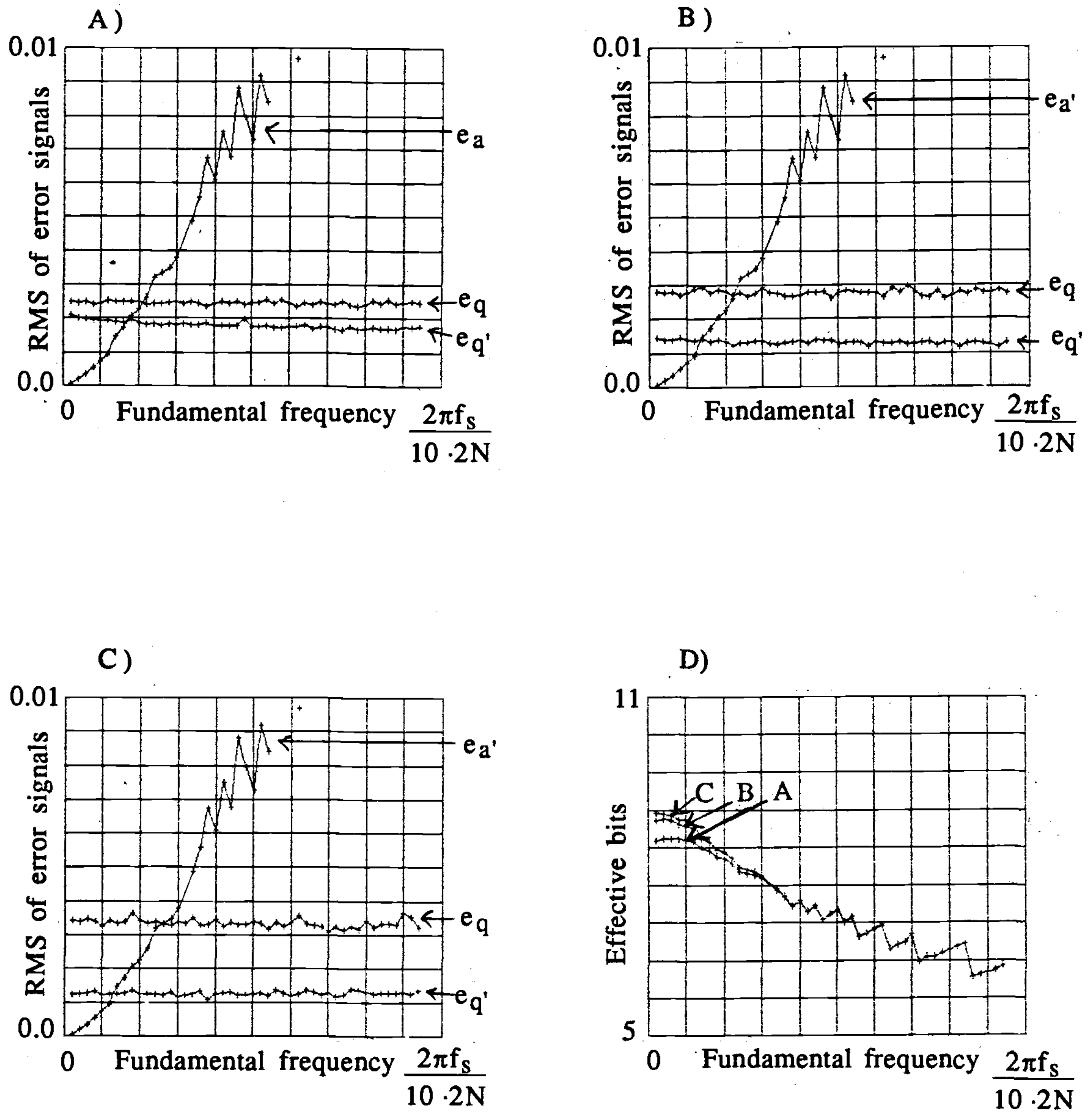


Figure 37

Simulated results showing effect of adding uniform, random noise to pulse train inputs having a pulse shape of $\tau = 0.05$ period. HERE, e_q and e_q' include added system noise. A MA filter is used on the original rate samples. Decimated output is inverse filtered to correct for MA filter passband roll-off. Decimation ration N = 10. A/D full scale = 2.1. Pulse amplitude range = 2.0.

A) Noise range is 0.0005·(pulse amp. range), ie. (~± 1/16 LSB of 8-bit A/D.)

B) Noise range is 0.002·(pulse amp. range), ie. (~± 1/4 LSB of 8-bit A/D.)

C) Noise range is 0.004·(pulse amp. range), ie. (~± 1/2 LSB of 8-bit A/D.)

D) Comparison of effective bits for methods in A), B), and C).

simulation results are summarized in figure 37. In these graphs, e_q, and e_q' include the added system noise[17] . Even though more noise is being added, the overall effect is that an optimal compromise is obtained when the amount of uniform noise added is about $\pm \frac{1}{2}$LSB (or perhaps slightly greater), as seen in figure 37(D).

It has been demonstrated by simulations that for sinusoidal inputs the resolution enhancement of the proposed lowpass system is within $\frac{1}{2}$ an effective bit of the theoretical results predicted in chapter 4. It is seen that for $10 < N \leq 500$ the correlation of the quantization noise produces a highpass spectrum and the resolution enhancement is higher than predicted. For $N > 500$ the correlation of the quantization noise produces a lowpass spectrum and the resolution enhancement decreases. Simulations have demonstrated that when a small dither signal is added to the input, the resolution enhancement is very close to that derived in chapter 4 using the white quantization noise model. This performance is within $\frac{1}{4}$ of an effective bit from that of an ideal lowpass system. For a class of periodic, non-bandlimited inputs it was demonstrated that the overall acquisition error, including alias error, is significantly reduced by the proposed lowpass system. For low fundamental frequencies it was shown that the operation of amplitude correction offers no substantial improvement in overall effective bits. The effects of dither for improving resolution of the lowpass system for squarewave-like signals was also demonstrated by simulations.

[17] Define $e1_q(n) = Q[x(n) + w(n)] - x(n)$, then the RMS of the total noise before processing is calculated,

$$e_q = \sqrt{\frac{1}{L}\sum_{n=0}^{L-1} e1_q^2(n)}$$

e_q' is calculated by passing $e1_q(n)$ through the system in figure 19, then calculating the RMS of the residual error.

Chapter 6

Conclusion

A computationally simple lowpass system has been proposed which enhances the average resolution of an A/D converter. The system meets all the objectives of this thesis, as stated in section 1.1:

1) It can operate at sample rates over a broad range of frequencies. The upper speed limit could reach hundreds of Megasamples/second due to the fact that addition is the most complex arithmetic operation required.

2) It is a low-cost modification to a raw, high-speed A/D converter. Primarily, the only extra component required is a high-speed accumulator. Amplitude correction can be processed off-line on low-speed computation facilities.

3) The absolute error of the raw quantizer is maintained for all inputs. The worst case is a constant-valued input. In this case, there would be no resolution enhancement. However, in actual system, ambient noise, or an added dither signal would ensure that there would always be some resolution enhancement.

This simple system is expected to provide about $\frac{1}{2}$ the resolution enhancement (in effective bits) as a substantially more complex noise shaping coder, when high speed operation is desired. Dither was shown to enhance the resolution enhancement for signals where the signal varies slowly for a large percentage of time. The system was also shown to be useful in reducing alias errors for non-bandlimited input signals.

References

[1] R.W. Adams, "Design and Implementation of an Audio 18-bit Analog-to-Digital Converter Using Oversampling Techniques," *J. Audio Eng. Soc.*, vol. 34, No. 3, pp. 153-166, Mar. 1986.

[2] B.P. Agrawal and K Shenoi, "Design Methodology for ΣΔM," *IEEE Trans. Comm.*, Vol. COM-31, No. 3. pp. 360-369, Mar. 1983.

[3] E.D. Banta, "On the Autocorrelation Function of Quantized Signal Plus Noise," *IEEE Trans. Info. Theory*, pp. 114-117, Jan. 1965.

[4] R.A. Belcher, "Apparatus and Methods for Analogue-to-Digital Conversion," U.S. Patent 4 621 254.

[5] B.E. Boser, K. Karmann, H. Martin, and B. Wooley, "Simulating and Testing Oversampled Analog-to-Digital Converters," *IEEE Trans. Comp. Aided Design*, Vol. 7, No. 6, pp. 668-674, June 1988.

[6] L. Richard Carley, "An oversampling Analog-to-Digital Converter Topology for High-Resolution Signal Acquisition Systems," *IEEE Trans. Circits Syst*, Vol. CAS-34, No. 1, pp. 83-90, Jan. 1987.

[7] T. Cataltepe, A.R. Kramer, L.E. Larson, G.C. Temes and R.H. Walden, "Digitally Corrected Multi-bit ΣΔ Data Converters," *Proc. IEEE Intnl. Symp. Circ. Syst.*, Portland, Oregon, May 8-11, 1989, pp. 647-650.

[8] T.A.C.M Claasen, W.F.G Mecklenbrauker, J.B.H Peek, and N. van Hurck, "Signal Processing Method for Improving the Dynamic Range of A/D and D/A Converters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 529-537, Oct. 1980.

[9] T.A.C.M Claasen and A. Jongepeir, "Model for the Power Spectral Density of Quantization Noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 914-917, Aug. 1981.

[10] C.C. Cutler, "Differential Quantization of Communication Signals," U.S. Patent 2 605 361, July 29, 1952, and U.S. Patent 2 724 740, Nov. 22, 1955.

[11] J.P. Deyst, Rough draft of M.I.T. M.S. Thesis written while the author was an intern at Tektronix, Inc., Beaverton, Ore. -- obtained from Shiv Balakrishnan of Tektronix, Aug. 3, 1988.

[12] P. Elias, "Predictive Coding," *IRE Trans. Info. Theory*, IT-1, No. 1, pp. 16-33, March 1955.

[13] IEEE Measurements and Analysis Committee of the IEEE Instrumentation and Measurement Society, "Standard for Waveform Recorders (Working Draft P1057/D8," 29 August, 1987.

[14] H. Inose, Y. Yasuda, and J. Murakami, "A telemetering System by Code Modulation -- $\Delta-\Sigma$ Modulation," *IRE Trans. Space Electron. Telemetry*, vol. SET-8, pp. 204-209, Sept. 1962.

[15] J.S Lim and A.V. Oppenheim, *Advanced Topics in Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, Chpt. 3, 1988.

[16] Y. Matsuya, et al., "A 16-bit Oversampling A-D Conversion Technology Using Triple Integration Noise Shaping," *IEEE J. Solid State Cir.*, Vol. SC-22, No. 6, pp. 921-929, Dec. 1987.

[17] J.H. McClellan, T.W. Parks, and L.R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," *IEEE Trans. Audio and Electroacoustics*, Vol. AU-21, No. 6, pp. 506-526, Dec. 1973.

[18] Motorola, Inc., "DSP56ADC16 16-Bit Sigma-Delta Analog-to-Digital Converter" (ADI1525 Technical Data), 1989.

[19] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[20] L. Schuchman, "Dither Signals and Their Effect on Quantization Noise," *IEEE Trans. Comm. Tech.*, vol. COM-12, pp. 162-165., Dec. 1964.

[21] B.S. Song, "A 10.7-MHz Switched-Capacitor Bandpass Filter," *IEEE J. Solid State Circuits*, vol. 24, No. 2, pp. 320-324, April 1989.

[22] H.A. Spang and P.M. Schultheiss, "Reduction of Quantization Noise by Use of Feedback," *IRE Trans. Comm. Syst.*, Vol. CS-10, pp. 373-380, Dec. 1962.

[23] A.B. Sripad and D.L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," *IEEE Trans. ASSP*, Vol. ASSP-25, pp. 442-448, Oct. 1977.

[24] S.K. Tewksbury and R.W. Hallock, "Oversampled Linear Predictive and Noise-Shaping Coders of Order N>1," *IEEE Trans. Circ. Syst.*, Vol. CAS-25, No. 7, pp. 436-447, July 1978.

[25] K. Uchimura, T. Hayashi, T. Kimura, and A. Iwata, "Oversampling A-to-D and D-to-A Converters with Multistage Noise Shaping Modulators," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, No. 12, Dec. 1988.

[26] B. Widrow, "Statistical Analysis of Amplitude Quantized Sampled Data Systems," *Trans. Amer. Inst. Elec. Eng. Pt. 2, Applications and Industry,* , vol. 79, pp. 555-568, Jan. 1960.

APPENDICES

Appendix 1:

Error Calculation Methods

This appendix defines many of the specific error discussed in this report and explains the methods used to calculate these errors from the simulations. First, some preliminary definitions are given, followed by definitions and calculation methods for specific errors.

A1.1 Preliminary definitions

Signal average power will be used extensively in the measurement of errors of the simulated results.[18]

For a discrete-time signal, average power will be defined by

$$U_x = \frac{1}{L} \sum_{n=0}^{L-1} x^2(nT) \qquad \text{(A1)}$$

and for continuous time

$$U_x = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} x^2(t)\ dt \qquad \text{(A2)}$$

[18] Other practical measures could be $\sum_{n=0}^{L-1} |e(nT)|) \leftrightarrow \int_{t_0}^{t_f} |e(t)|\ dt$

or $\max_{n\in[0,L-1]\ \text{(integers)}}(e(nT)) \leftrightarrow \max_{t\in[t_0,t_f]}(e(t))$

where L is the number of samples in the record covering the time interval $[t_f, t_0]$ and T is the sample period.

The discrete Fourier transform (DFT) will be defined

$$X(e^{j\frac{2\pi}{L}k}) = \sum_{n=0}^{L-1} x(n)\, e^{-j\frac{2\pi}{L}n\,k} \qquad \text{(A3)}$$

The discrete Fourier series (DFS) is identical to the DFT with the provision that the record length is the period of the sequence. The DFS will be denoted by X(k) to distinguish it from the DFT.

A1.2 Method for measuring passband attenuation error

This error is easiest to define and calculate in the frequency domain. The passband attenuation error spectrum is defined

$$E_p(e^{j\omega}) = \begin{cases} \left(1-H_{MA}(e^{j\omega})\right) X(e^{j\omega}) \ , \ |\omega| < \frac{\pi}{N} \\ 0 \ , \ \text{otherwise} \end{cases} \qquad \text{(A4)}$$

where $X(e^{j\omega})$ is the spectrum of the unquantized discrete-time input calculated over the record of L samples, that is,

$$X(e^{j\omega}) = \sum_{n=0}^{L-1} x(n)\, e^{-j\omega n} \qquad \text{(A5)}$$

For all the cases that were simulated, the input signals were periodic and had line spectra. Therefore, the average power of the passband attenuation error can be exactly calculated using the DFS. Using Parseval's relation

$$\sum_{n=0}^{L-1} |x(n)|^2 = \frac{1}{L}\sum_{k=0}^{L-1} |X(k)|^2 \tag{A6}$$

the average power can be calculated by

$$U_p = \frac{1}{L^2}\left(|E_p(0)|^2 + 2\sum_{k=1}^{b} |E_p(k)|^2 \right) \tag{A7}$$

where b is the largest integer which satisfies $\frac{2\pi}{L} b \leq \frac{\pi}{N}$. It is assumed in (A7) that the input signal is real and therefore has a conjugate symmetric spectrum. If x(n) is decimated by N, then U_p is divided by N (see equation (A15)).

A1.3 Method for measuring alias errors

First, a general framework for measuring alias errors is given. Then a specific method, used in conjunction with the simulations, is presented.

A1.3.1 General method for measuring alias error after decimation

We would like to keep all measurements using discrete-time since our original input will be in the form of a time series, as opposed to interpolating increasingly fine to approximate continuous time. Another problem with using continuous time is that we would have to be careful that all errors in interpolation and numerical integration would not become significant. The concept of aliasing, however, makes more sense in continuous time. To resolve this issue the meaning of aliasing will be investigated. Aliasing is usually discussed in conjunction with the sampling theorem. If the original continuous-time input signal is bandlimited, ie. $|X(\omega)| = 0$, for $|\omega| > \omega_B$, and is sampled at a rate greater than the Nyquist rate, then if the sampled signal is decimated by $N < \pi/\omega_B$, theoretically, the original continuous-time signal could still be reconstructed. But if we decimate by greater than π/ω_B, then it would be impossible to reconstruct the signal in general. Note that the process of decimation does not introduce error in the samples, it only introduces error in the reconstructed continuous time signal. Thus we cannot directly "see" aliasing error using discrete time. But we can measure it indirectly; we can measure the difference between a decimated sequence from which perfect reconstruction could be performed, and a decimated sequence from which aliasing would occur. For example, suppose we have a bandlimited continuous-time signal x(t) which we sample with infinite amplitude precision, at slightly greater than the Nyquist rate. This forms the sequence x(nT). Now we decimate by N. Before we decimate, we pass the same sequence x(nT) through different discrete-time (digital) filters h_1 and h_{MA} (see figure 38) to produce y_1 and y_{MA}, respectively. Then we decimate by N to produce $y_1(nTN)$ and $y_{MA}(nTN)$. Using a sin(x)/x interpolator we produce the continuous-time "reconstructions" $y_1(t)$ and $y_{MA}(t)$. The average power of the alias error signal would be

$$U_a = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} [y_1(t) - y_{MA}(t)]^2 \, dt \qquad \text{(A8)}$$

using the continuous-time average power definition, and similarly

$$U_a = \frac{1}{L} \sum_{n=0}^{L-1} [y_1(nTN) - y_{MA}(nTN)]^2 = \frac{1}{L} \sum_{n=0}^{L-1} e_a{}^2(nTN) \qquad \text{(A9)}$$

using the discrete-time average power definition. Equation (A9) describes the general method used to measure alias average power for discrete-time signals--we pass the input sequence through a reference lowpass filter, which is perfectly, or almost perfectly bandlimited, similar to h_1, then pass the input sequence through another filter which is being evaluated and compare the two results. Note that h_1 and h_{MA} have the same magnitude

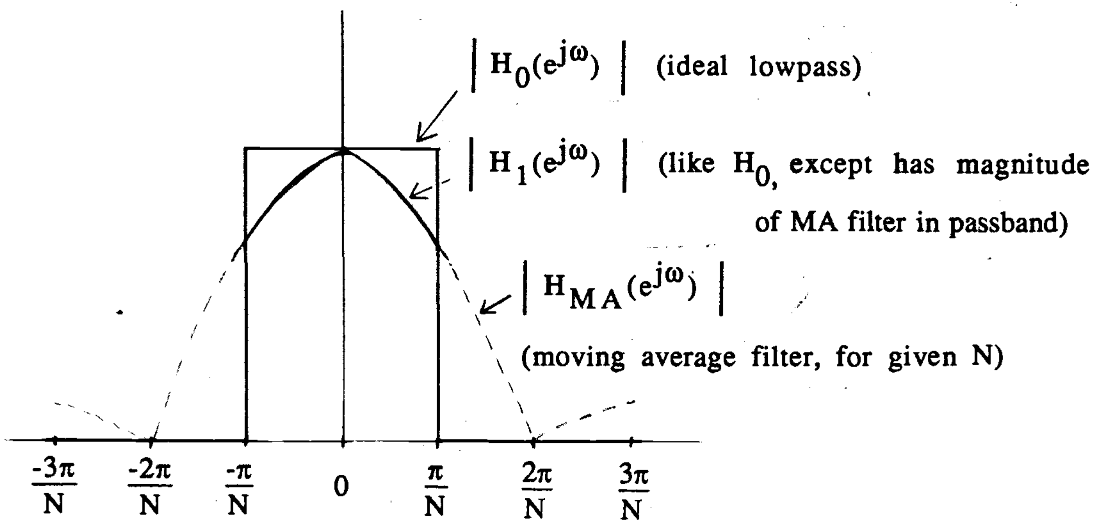

$|H_0(e^{j\omega})|$ (ideal lowpass)
$|H_1(e^{j\omega})|$ (like H_0, except has magnitude of MA filter in passband)
$|H_{MA}(e^{j\omega})|$
(moving average filter, for given N)
$\frac{-3\pi}{N}$
$\frac{-2\pi}{N}$
$\frac{-\pi}{N}$
0
$\frac{\pi}{N}$
$\frac{2\pi}{N}$
$\frac{3\pi}{N}$

Figure 38

Family of filters used in the analysis of alias and passband errors (no periodicity requirement on input).

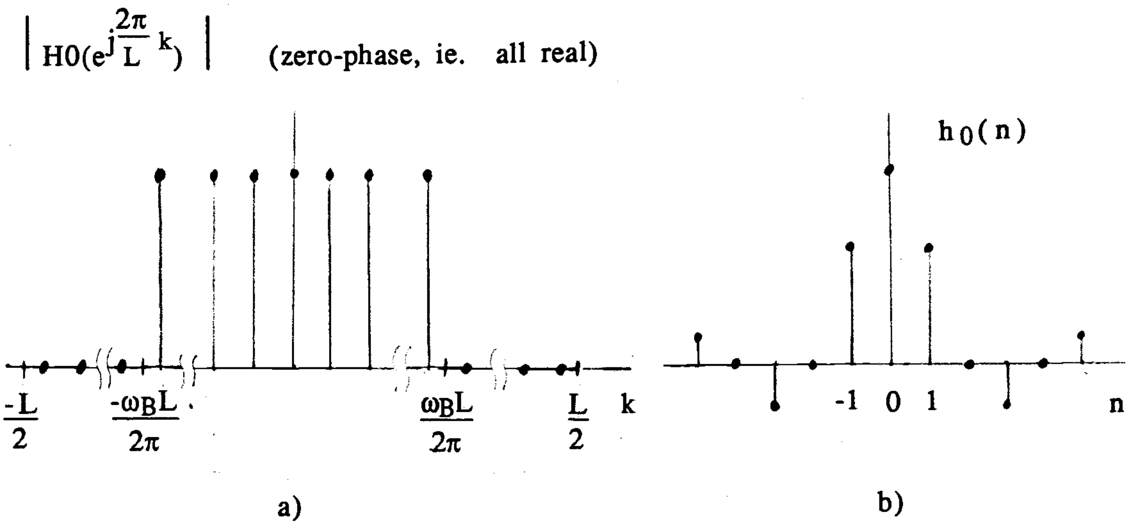

$|H0(e^{j\frac{2\pi}{L}k})|$ (zero-phase, ie. all real)
$\frac{-L}{2}$
$\frac{-\omega_B L}{2\pi}$
$\frac{\omega_B L}{2\pi}$
$\frac{L}{2}$
k
a)
$h_0(n)$
-1 0 1
n
b)

Figure 39

Family of filters used in analysis of alias and passband errors (periodic inputs).

a) Frequency response.
b) Impulse response (included to show time-domain correspondence of

frequency response in the passband. Thus, passband attenuation error is eliminated.

A1.3.2 Specific method for measuring alias error after decimation

This methods operates on the unquantized input sequence. In the model of quantization, ie. $Q[x(n)] = x(n) + e_q(n)$, we have separated the source of quantization error, $e_q(n)$, from the signal, $x(n)$. Alias and passband errors are only meaningfully defined over $x(n)$, therefore, $e_q(n)$ is not considered when analyzing alias effects.[19] This method only works accurately for periodic sequences. It works for any down sampling factor N. All analysis is performed in the frequency domain.

Assume we have a record of length L of a periodic sequence, x(n), which contains an exact number of periods. That is, $x(n) = x(n+\frac{L}{m})$ for $m \in \{1,2,...\}$. The DFT of this record is precisely the DFS of the periodic sequence. With these assumptions, we can realize a lowpass filter which has the same effect as an ideal lowpass filter. The easiest way is to multiply $X(e^{j\frac{2\pi}{L}k})$ by $H_0(e^{j\frac{2\pi}{L}k})$, where

$$H_0(e^{j\frac{2\pi}{L}k}) = \begin{cases} 1, & 0 \le k < \frac{\omega_B L}{2\pi} \text{ and } L - \frac{\omega_B L}{2\pi} < k \le L-1 \\ 0, & \frac{\omega_B L}{2\pi} \le k \le L- \frac{\omega_B L}{2\pi} \end{cases} \tag{A10}$$

$$0 \le \frac{2\pi k}{L} \le 2\pi$$

[19] See equation (A21).

Note that $H_0(k)$ is real. Thus, the corresponding impulse response $h_0(n)$ is is symmetrical about n=0 (zero-phase) (see figure 39).

The filter we want to analyze, the MA filter, is defined

$$h_{MA}(n - \frac{N-1}{2}) = \begin{cases} \frac{1}{N}, & 0 \leq n < N \\ 0, & \text{otherwise} \end{cases} \qquad \text{(A11)}$$

The time offset of $\frac{N-1}{2}$ is included so that the filter is zero-phase.

The DFS of the reference sequence $y_1(n)$ is obtained by multiplying the DFT of $H_1(e^{j\frac{2\pi}{L}k})$ and $X(e^{j\frac{2\pi}{L}k})$. The reference filter we use, $H_1(e^{j\frac{2\pi}{L}k})$, must have the same pass band magnitude response as the filter we are analyzing. So,

$$H1(e^{j\frac{2\pi}{L}k}) = H0(e^{j\frac{2\pi}{L}k})\ HMA(e^{j\frac{2\pi}{L}k}) \qquad \text{(A12)}$$

This is equivalent to circular convolution in the time domain. The output of the circular convolution represents one period of $y_1(n)$. Next, one period of the convolution of x(n) and $h_{MA}(n)$, (ie. $y_{MA}(n)$) must be obtained to compare with the reference $y_1(n)$. The implementation of circularly convolving $h_{MA}(n)$ and x(n) was performed by linearly convolving $h_{MA}(n)$ with x'(n), then removing transients, where

$$x'(n) = \begin{cases} x(n), & 0 \leq n < L+N-1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{(A13)}$$

One period of x(n), extended by N-1 samples

Thus,

$$y_{MA}(n) = \sum_{i=0}^{L-1} h_{MA}(k)\ x'(n-k) \qquad \frac{N-1}{2} \le n < L + \frac{N-1}{2} \qquad \text{(A14)}$$

Now we consider the effect of decimation on both $y_{MA}(n)$ and $y_1(n)$. We do not explicitly decimate in the time domain for $y_1(n)$. Instead we use the following formula which represents the effects of decimation in the frequency domain (this is exact for periodic, sequences, or for infinitely long records of any sequence) [15],

$$\mathrm{DFT}[y1(nN)] = \frac{1}{N}\sum_{i=0}^{N-1} Y1(e^{j2\pi(\frac{k}{LN} - \frac{i}{N})})$$

$$= \frac{1}{N}\sum_{i=0}^{N-1} X(e^{j2\pi(\frac{k}{LN} - \frac{i}{N})})\ H1(e^{j2\pi(\frac{k}{LN} - \frac{i}{N})})$$

$$= \frac{1}{N}\ X(e^{j2\pi\frac{k}{LN}})\ H1(e^{j2\pi\frac{k}{LN}}) \qquad \text{(A15)}$$

We do explicitly decimate $y_{MA}(n)$ in the time domain, thus by using (A6) we can write the final expression for alias error average power,

$$U_a = \frac{1}{L^2}\left(|E_a(0)|^2 + 2\sum_{k=1}^{b} |E_a(k)|^2 \right) \qquad \text{(A16)}$$

where b is the largest integer which satisfies $\frac{2\pi}{L} b \le \frac{\pi}{N}$, and

$$E_a(k) = \frac{1}{N}\ X(e^{j2\pi\frac{k}{L}})\ H1(e^{j2\pi\frac{k}{L}}) - \mathrm{DFT}[y_{MA}(nN)]\ e^{-j2\pi\frac{k}{LN}\frac{(N-1)}{2}} \qquad \text{(A17)}$$

Note that the phase modification to $\mathrm{DFT}[y_{MA}(nN)]$ in equation (A17) accounts for the fact that only non-transients are used.[20]

[20] See equations (A13) and (A14).

After inverse filtering is performed to provide amplitude correction we have,

$$U_{a'} = \frac{1}{L^2}\left(|E_a'(0)|^2 + 2\sum_{k=1}^{b} |E_a'(k)|^2 \right) \tag{A18}$$

where,

$$E_a'(k) = E_a(k) \frac{1}{H_{inv}(e^{j2\pi\frac{k}{L}})} \tag{A19}$$

For this method to be precise L/N must divide exactly, so that the length L/N DFT of the down sampled periodic sequence $y_{MA}(nN)$ is precisely the DFS (no spectral leakage).

A1.4 Method for Analyzing Quantization Error after Filtering

Throughout this study, quantization error is modeled as

$$Q[x(n)] = x(n) + e_q(n) \tag{A20}$$

where Q[·] represents the quantization operation. If we pass Q[x(n)] through the filtering/decimation system in figure 10b, we see that the output can be calculated as,

$$y_{MA}(nN) = \sum_{k=0}^{N-1} h_{MA}(k)\, x(nN-k) + \sum_{k=0}^{N-1} h_{MA}(k)\, e_q(nN-k) \tag{A21}$$

The first sum determines e_p and e_a, and the second sum determines e_q'. Equation (A21) clarifies why the passband and alias error were calculated on the unquantized signal; the different error calculations separate nicely with this approach.

The average power of the filtered quantization error sequence can be expressed,

$$
\begin{aligned}
U_{q}' &= \frac{1}{L}\sum_{nN=0}^{L-1} e_{q}'^{2}(nN) \\
&= \frac{1}{L}\sum_{nN=0}^{L-1}\left(\sum_{k=0}^{N-1} h_{MA}(k)\; e_{q}(nN-k)\right)^{2} \qquad \text{(A22)}
\end{aligned}
$$

If the output precision of the decimator is less than the number of bits in the maximum sum of $y_{MA}(nN) = \sum_{k=0}^{N-1} h_{MA}(k)\; Q[x(nN-k)]$, a second quantization stage is realized. This stage represents rounding the result to a specified number of bits,

$$
\begin{aligned}
e_{q}''(nN) = \mathrm{RND}&\left(\mathrm{NORM}\left(\sum_{k=0}^{N-1} h_{MA}(k)\; Q[x(nN-k)]\right)\right) - \\
&\mathrm{NORM}\left(\sum_{k=0}^{N-1} h_{MA}(k)\; Q[x(nN-k)]\right) \qquad \text{(A23)}
\end{aligned}
$$

where $e_{q}''(nN)$ is the error sequence caused by rounding. NORM stands for a normalization operation where the accumulation of the current filter output is right-shifted such that the longest possible accumulation would not overflow the desired precision MSB. RND stands for the rounding operation where the least significant bits of the normalized accumulation are rounded off.

$e_{q}'(nN)$ can be updated to input the rounding error $e_{q}''(nN)$

(total with rounding)   (total without rounding)   (rounding)

$$e_q'(nN) \leftarrow e_q'(nN) + e_q''(nN) \qquad (A24)$$

and,

$$U_q' = \frac{1}{L}\sum_{nN=0}^{L-1} e_q'^2(nN) \qquad (A25)$$

assuming $e_q'(n)$ (no rounding) is uncorrelated with $e_q''(n)$.

If the output sequence is to be amplitude corrected, this operation must be performed on the error sequence as well. So, finally,

$$e_q'(nN) \leftarrow A[e_q'(nN)] \qquad (A26)$$

where $A[\cdot]$ stands for the amplitude correction operation (scaling or inverse filtering), and again

$$U_q' = \frac{1}{L}\sum_{nN=0}^{L-1} e_q'^2(nN) \qquad (A27)$$

Appendix 2:

A Broader Definition of Effective Bits

The definition of effective bits is [13],

$$EB = B - \log_2\left(\frac{\text{actual rms error}}{\text{ideal rms error}}\right) = B - \frac{1}{2}\log_2\left(\frac{\text{actual ms error}}{\text{ideal ms error}}\right) \qquad \text{(A28)}$$

where "ideal rms error" = $\frac{q}{\sqrt{12}}$ represents the rms error of a B bit quantizer assuming the white uniform noise model. The "actual rms error" is the rms of the difference between the A/D output samples and a signal which is a least squares fit of the samples to a sinusoid. Three comparable definitions of EB are also used in this thesis.

Theoretical analysis: Theoretical variances of stochastic processes are substituted for the "actual ms error" term. In this case,

$$EB = B - \frac{1}{2}\log_2\left(\frac{\sigma_{e_q}^{\ 2}}{\frac{q^2}{12}}\right) \qquad \text{(A29)}$$

where $\sigma_{e_q}^{\ 2}$ is the theoretical variance of the quantization noise. When white system noise, w(n), is added to the input,

$$EB = B - \frac{1}{2}\log_2\left(\frac{\sigma_{e_q}^{\ 2} + \sigma_w^{\ 2}}{\frac{q^2}{12}}\right) \qquad \text{(A30)}$$

where it is assumed that $e_q(n)$ and w(n) are uncorrelated, so variance adds.

Empirical: Average power measures, different from those in the original definition of effective bits, are substituted for "actual ms error." In the method for analyzing quantization error described in appendix A1.4, the average power of $e_q'(n)$, U_q', is substituted for "actual ms error",

$$EB = B - \frac{1}{2}\log_2\left(\frac{U_q'}{\frac{q^2}{12}}\right) \tag{A31}$$

In the simulation results presented in section 5.2, average powers from different sources are summed and are substituted for "actual ms error." For example,

$$EB = B - \frac{1}{2}\log_2\left(\frac{U_q' + U_a' + U_p}{\frac{q^2}{12}}\right) \tag{A32}$$

It is assumed that the different errors are not cross correlated, so average power adds.

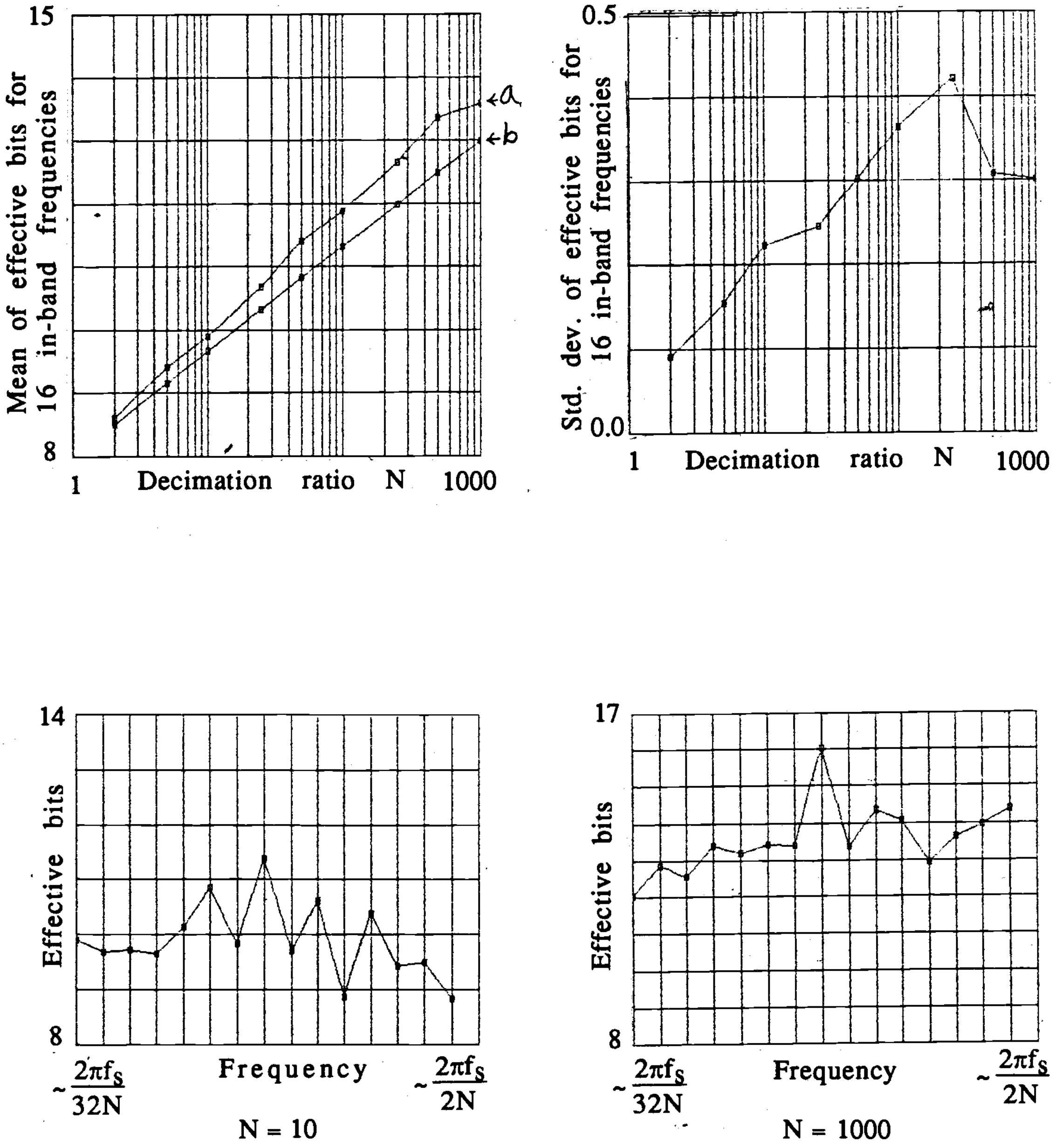
Appendix 3:

When the sampling frequency is an exact multiple of the input signal frequency

One set of simulations was performed where the sampling frequency was an exact multiple of the input frequency--a rare case in actual sampling ("fraction" = 0 in equation (33)). These simulation results are presented in figure 40. These results were inconsistent with the other simulations where

fraction = $\frac{37}{87}$

and fraction = $\frac{11}{173}$.



a - Simulated results using MA decimation filter.
b - Theoretical results using an ideal lowpass filter and white quantization noise models.

Figure 40

Typical effective bits for the unusual case where the sampling frequency is an exact multiple of the input sinusoid frequency. Decimated output is scaled to correct for MA filter roll-off.