THE UNIVERSITY OF CALGARY

Analysis, Characterization, and Design of a Class of Oversampled Sigma-Delta Converters

by

Thomas P. Borsodi

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

CALGARY, ALBERTA

September, 1995

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled, "Analysis, Characterization, and Design of a Class of Oversampled Sigma-Delta Converters", submitted by Thomas P. Borsodi in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE.

B. Nowrouzian

Dr. B. Nowrouzian, Supervisor & Chairman
Dept. of Elect. & Comp. Engineering

L. T. Bruton

Dr. L. T. Bruton
Dept. of Elect. & Comp. Engineering

Ed Nowicki

Dr. E. Nowicki
Dept. of Elect. & Comp. Engineering

J. A. R. Blais

Dr. J. A. R. Blais
Dept. of Geomatics Engineering

Date: September 25, 1995

ABSTRACT

This thesis presents the analysis, characterization, and design of a class of oversampled sigma-delta converters. An overview of sigma-delta conversion, modern uses of sigma-delta converters, and important issues regarding these converters is presented. Then, the closed-form solution for the granular quantization error for this class of sigma-delta converters is derived using a state-space approach. An input signal bound to guarantee the quantizer is not overloaded is also derived. A closed-form solution for the quantizer output signal based on state-space equations is derived along with an open-loop equivalent system based on the closed-form solution for the granular quantization error. Stability issues and spectral analysis methods of sigma-delta converters are examined. Finally, design techniques for cascaded sigma-delta converters are presented. New converters are developed and analysed. The operation of these new converters is then characterized with regard to signal-to-quantization noise ratio through simulation.

Acknowledgement

I would like to thank my supervisor, Dr. B. Nowrouzian, for his guidance and encouragement during the course of this research, and for his criticism and guidance during the preparation of this thesis.

I would also like to thank the Alberta Microelectronic Centre for their financial support in the form of a graduate student scholarship.

In addition, I would also like to acknowledge the support provided by Micronet Network of Centres of Excellence, and the financial support provided by Bell Northern Research.

Finally, I would also like to acknowledge the efforts of Mr. A. Borsodi in the proofreading of this work and for his useful and lively discussions in general that helped me to maintain a certain amount of calmness during these past months.

To my parents,

who always encouraged their

children to strive for knowledge.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

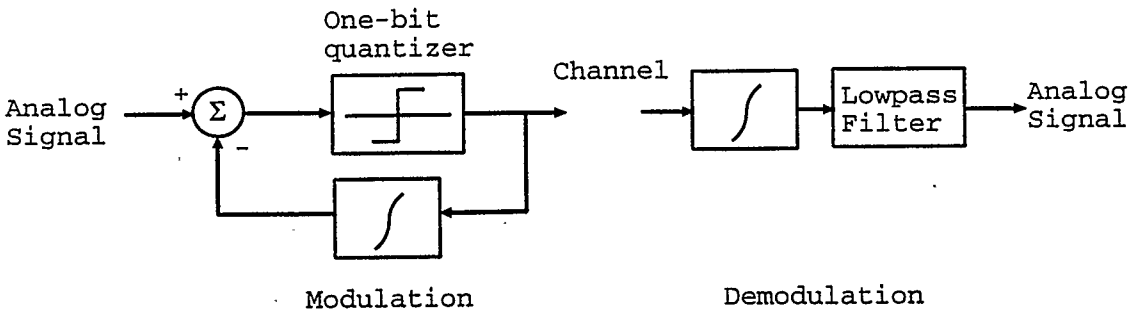
CHAPTER 1

INTRODUCTION

1.1 Historical background of sigma-delta ($\Sigma - \Delta$) modulation

The widely recognized field of sigma-delta ($\Sigma - \Delta$) conversion has its origins back in the early 1960's. It was at that time when H. Inose, Y. Yasuda, and J. Murakami [IYM62] with the Faculty of Engineering at the University of Tokyo developed the delta-sigma ($\Delta - \Sigma$) conversion technique as a code modulation scheme in the field of communications. They proposed the new scheme as an alternative to the conventional delta modulation where pulses are generated by the differentiation of the amplitude of the input signal in an encoder at the transmit end, and integrated at the decoder on the receive end to obtain the original waveform. This technique is known to suffer from the effects of transmission noise having an accumulative error upon the demodulated signal. The $\Sigma - \Delta$ modulator provides a solution to this drawback, involving the integration of the input signal prior to modulation so that the generated output pulses would carry the information corresponding to the input signal amplitude (see Fig.1.1).

In recent years, the fields of signal processing and communications have become more important with the continuing advances in very large scale integrated (VLSI) circuit technology. This is because the robustness, accuracy, and flexibility of VLSI technology have created new areas for signal processing applications and new implementation alternatives. One such area that is receiving increased attention is that



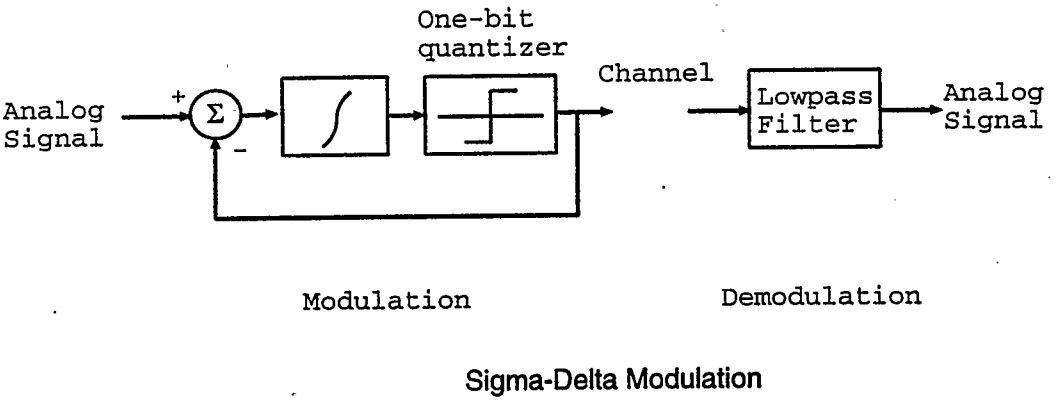


Figure 1.1. Comparison of delta and sigma-delta modulation

of analog-to-digital (A/D) and digital-to-analog (D/A) conversion. This is mainly due to the fact that continuing improvements in digital signal processing resolution are causing the resolution requirements for interface circuits (*i.e.* A/D and D/A converters) to increase as well.

There are several methods available for accurate A/D conversion. These include:

a) Parallel (*also known as* flash) as well as serial-parallel converters.

b) Pipelining and multiplexing converters.

c) Serial (*also known as* successive-approximation) converters.

d) Counting converters.

e) Oversampling converters.

A thorough discussion of each of these conversion methods may be found in the existing literature such as [LT93]. These methods may also be used for D/A conversion also. In order to facilitate the discussion of the increased resolution requirements, it is expedient to discuss some of the drawbacks that limit the resolution capability of the above named A/D conversion methods.

Parallel Converters These converters offer rapid conversion in one clock cycle with the use of $2^N + 1$ comparators (where N is the desired number of bits). This creates problems for resolution exceeding 9 bits, as the circuit complexity, power dissipation, and chip area all become large. In addition, the accuracy of the comparisons increases for a larger number of bits, thus making the necessary manufacturing tolerances of circuit components more difficult to achieve.

Pipelining and multiplexing converters Pipelining converters are much more expensive in chip area compared to parallel converters as they require N sample-and-hold stages in addition to $N-1$ analog gain stages. The multiplexing converters employ N sample-and-hold and converter stages with one N-bit multiplexor to achieve an A/D conversion in M clock cycles. Additionally, skew or timing jitter in any of the N clocks will translate into more accumulated noise in the system.

Serial converters These converters require very few analog components, but require N clock cycles to process all the bits of an N-bit conversion. The complexity of the digital logic and storage, however, becomes much higher than previously mentioned methods.

Counting converters These converters like the serial converters mentioned above contain only a modest amount of components such as multiple converters, digital logic, a counter, and a D/A converter. However, these converters trade a larger processing time for the desired accuracy, as they require 2^N steps to achieve an N-bit conversion.

Oversampling converters It has been clearly demonstrated in the literature that the use of a coarse quantizer embedded in a feedback loop and operated at a sampling rate much higher than the *Nyquist* rate (hence the term *oversampled*), will result in a high resolution digital approximation of the original analog input signal after processing the output digital bit stream. One apparent drawback is that for high resolution, the oversampling ratio (OSR) which is the ratio of the sampling frequency f_s over twice the highest input signal frequency f_o must be very large. A second drawback is that due to the simple structure of this network,

there exists a correlation between the input signal and the generated quantizer error at DC and very low frequencies which may cause resonance effects. This in turn causes large inband tones to be generated in the quantization error (seriously degrading resolution). These problems may be overcome with the use of more complex networks containing more loops, which will be discussed later in Section 1.3.

One of the inherent benefits of oversampled $\Sigma-\Delta$ A/D converters is that the anti-aliasing requirement for the circuit becomes less stringent. The sampling of analog signals has been extensively established, and is governed by the minimum rate at which a signal must be sampled to prevent loss of information (the *Nyquist* rate). However, no restrictions have been placed on the upper bound at which the signal may be sampled except those due to technological limitations. For example, in digital audio where the passband is limited to 20 kHz, the designated sampling frequency is 44.1 kHz. This would require an anti-aliasing filter with a sharp roll-off (see Fig. 1.2). Note, that in Figs. 1.2 and 1.3, the frequency axis has been exaggerated to show the change in the anti-aliasing filter roll-off more clearly. Utilizing the oversampling

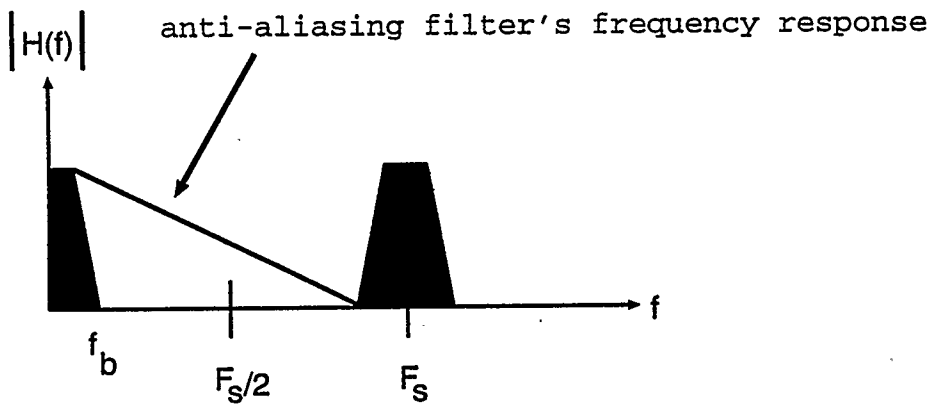


Figure 1.2. Anti-aliasing filter requirements

techniques however, causes the signal spectra to become more further separated in frequency due to the higher sampling rate, thus reducing the requirements for the anti-aliasing filter (see Fig. 1.3). In some special applications, with a sufficiently high oversampling ratio, the anti-aliasing requirements for certain circuits may be met with a simple RC filter.

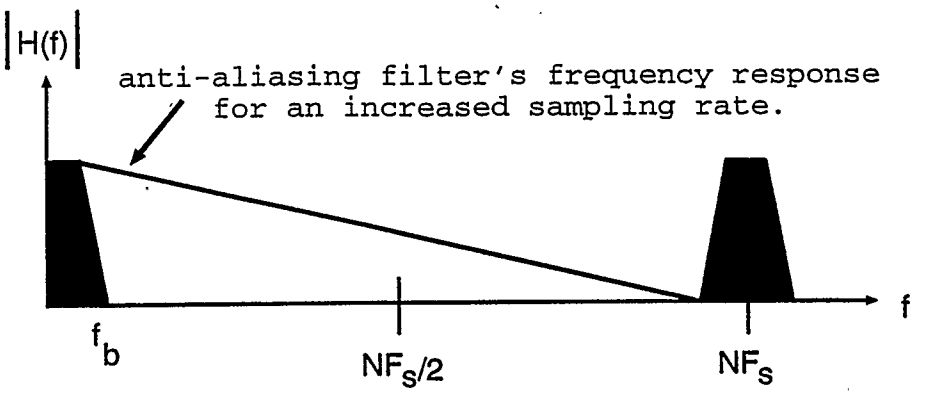


Figure 1.3. Reduced anti-aliasing requirements

Originally developed for code modulation in the field of communications, the $\Sigma-\Delta$ modulation technology has now evolved dramatically in the past two decades. This technology is currently employed in a large number of consumer audio products such as Compact Disc players, digital tape decks and other electronic stereo products. It is also employed in the digital conversion of intermediate frequency (IF) in digital radio products, interface circuits for instrumentation, and is being utilized quite heavily in telecommunication codecs.

1.2 Sigma-Delta converter operation

This section is concerned with a brief study of the basic operating principles of $\Sigma - \Delta$ converters. For simplicity, let us discuss $\Sigma - \Delta$ converter operation for a single-loop converter (see Fig. 1.4). This converter is implemented with a differential

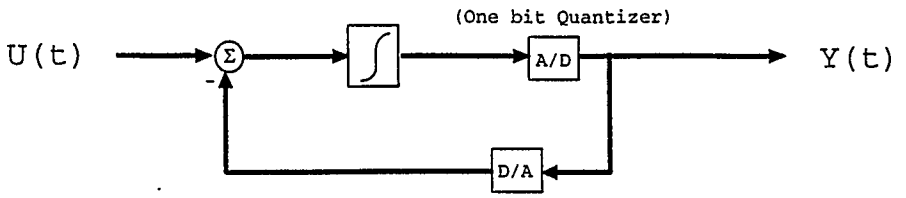


Figure 1.4. Single-loop sigma-delta converter

integrator, a coarse quantizer (a simple comparator for 1 bit quantization), a D/A converter, and a delay. The input to the circuit is fed to the quantizer via the integrator, after which the resulting signal is fed back and subtracted from the input signal. Thus the slowly moving short term average value of the quantized signal tracks the average input due to the forcing action of the feedback signal. These structures, also known as *noise shapers*, have the additional property of reducing the quantization error spectral density at low frequencies and increasing the quantization error spectral density at higher frequencies. This noise shaping property is subsequently increased for converters containing multiple loops or stages. In order to facilitate the discussions in this and the following chapters, it is more useful to represent the above sampled data system with its discrete-time equivalent representation shown in Fig. 1.5. Although the operation of such oversampled systems may seem straightforward,

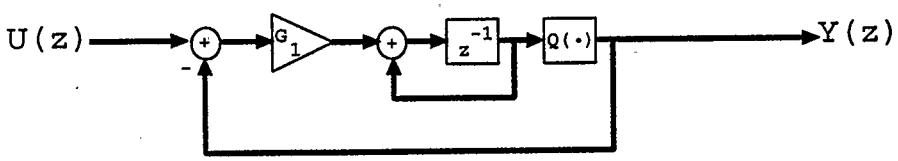


Figure 1.5. Discrete-time equivalent system

there are important issues regarding the analysis, characterization, and design of such systems. Analysis is a key issue because exact analysis methods for the quantization error will provide essential insight into the basic nature of such devices that approxi-

mation methods for the quantization error cannot predict. Next, characterization of system output behaviour for various input signals is important as it is the "yardstick" by which implemented circuits are judged as to their performance capabilities and therefore more exacting analysis techniques that characterize system behaviour must be used to overcome the shortcomings of analytical approximations. Finally, design is a key issue that requires some rigourous methods so that implemented circuits actually meet the original design criteria and not the designer's perceived ideal output response.

1.3 Survey of design methods for lowpass $\Sigma - \Delta$ converters

There are different methods available for the design of oversampled $\Sigma - \Delta$ converters. These converters may be used for either lowpass or bandpass signal applications such as A/D conversion of low frequency audio signals in CD players or A/D conversion of high frequency IF frequency signals in digital radio. This section will survey design methods used for lowpass applications, and the next section will survey design methods used for bandpass applications. As mentioned in Section 1.1, simple oversampled converters suffer from two shortcomings, the first being that an extremely large oversampling ratio must be used to reduce the in-band quantization error and obtain high resolution, and the second being that the quantizer error signal is correlated to the input signal for certain inputs that are not sufficiently random. These problems have been circumvented by using multibit quantization, multiloop configurations, or cascaded simple multiloop networks. It was noticed quite early in this field of research, that further decorrelation between the input signal and the quantizer error occurs for networks employing more than a single loop in a $\Sigma - \Delta$ converter [Can85] and that such structures shaped more low frequency noise into the higher frequency bands for the same oversampling ratios (see Fig. 1.6 for a comparison).

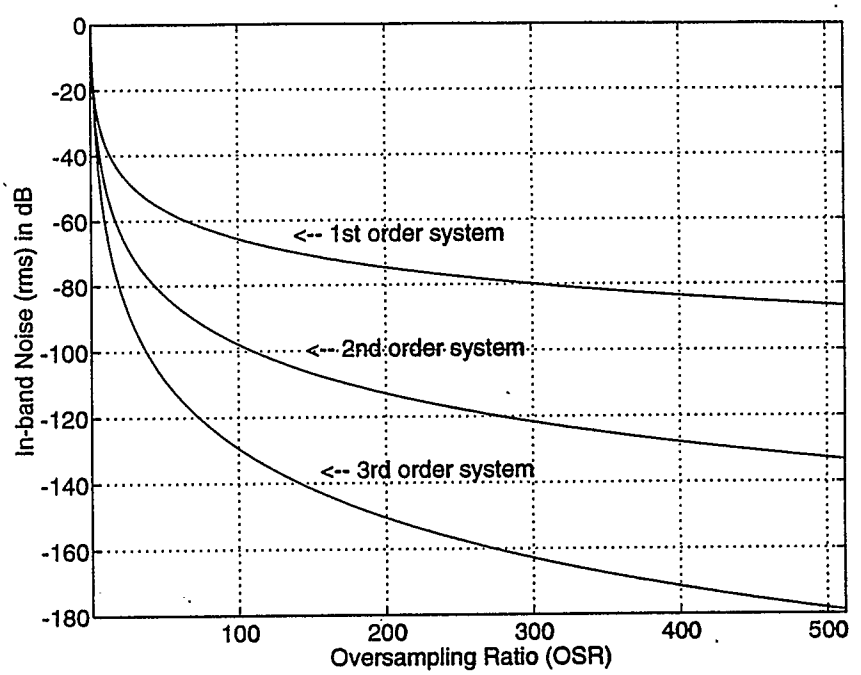


Figure 1.6. In-band noise versus oversampling ratio comparison for higher-order converters

From the above research, two distinct approaches emerged for the design of higher-order converters providing improved noise shaping. These approaches are discussed in the following two subsections.

1.3.1 Multiloop $\Sigma - \Delta$ converters

Research into higher-order oversampled converters [Can85] made it clear that improved resolution becomes possible by using circuit configurations with higher-order noise shaping properties to achieve a much higher signal-to-quantization noise ratio for the same given modest oversampling ratio (the benefits are not apparent for low oversampling ratios). However, it was also determined that for coarse quantization (1 bit), system stability became an important issue for structures with $N \geq 3$ loops, unless multibit quantization was used [CT92].

In general, multiloop $\Sigma - \Delta$ converter circuits may be represented by using the configuration in Fig. 1.7. From this representation, two transfer functions can be de-

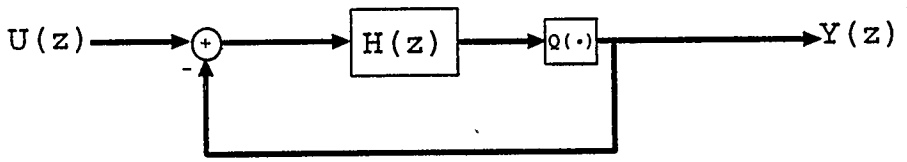


Figure 1.7. The general multiloop converter representation

rived. The first is between the input signal and the output, which is referred to as the signal transfer function (STF), and the second is between the generated quantization error signal and the output and is referred to as the noise transfer function (NTF). These transfer functions may be represented as functions of the feedforward transfer

function $H(z)$ in accordance with

$$\text{STF:} \quad H_S(z) = \frac{H(z)}{1 + H(z)} \tag{1.1}$$

and

$$\text{NTF:} \quad H_N(z) = \frac{1}{1 + H(z)} \tag{1.2}$$

Thus, for circuit configurations with a given feedforward transfer function $H(z)$, if care is not taken in choosing the transfer function parameters, circuit instability will result. In the design of higher-order single quantizer $\Sigma - \Delta$ converters, the following circuit configurations have been used with varying amounts of success.

1) A chain of integrators configuration with weighted feedforward summation.

2) A chain of integrators configuration with distributed feedback.

3) A chain of integrators configuration with distributed feedback and distributed feedforward paths.

4) A chain of integrators configuration with distributed feedback and local resonator feedbacks.

5) A chain of integrators configuration with distributed feedback and distributed feedforward paths with local resonator feedbacks.

The first three methods may be attributed to the work of Chao, Naddeem, Lee, and Sodini [CNLS90] who implemented an Nth order loop subsystem with feedforward and feedback coefficients as shown in Fig. 1.8, resulting in the following signal and noise transfer functions.

$$H_S(z) = \frac{\sum_{i=0}^{N} A_i(z-1)^{N-i}}{z[(z-1)^N - \sum_{i=1}^{N} B_i(z-1)^{N-i}] + \sum_{i=0}^{N} A_i(z-1)^{N-i}}$$

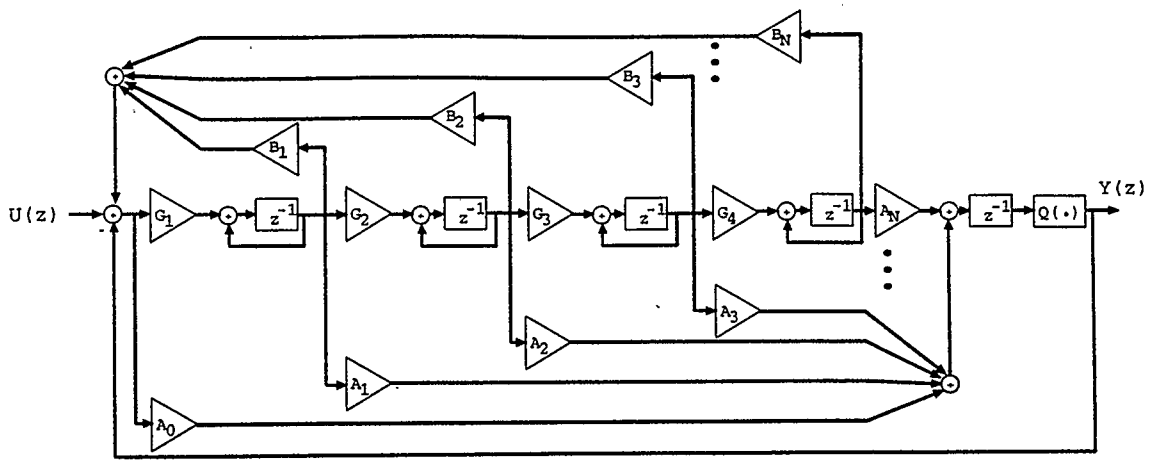


Figure 1.8. The Nth order loop topology with feedforward and feedback coefficients

$$H_N(z) = \frac{(z-1)^N - \sum_{i=1}^{N} B_i(z-1)^{N-i}}{z[(z-1)^N - \sum_{i=1}^{N}(z-1)^{N-i}] + \sum_{i=o}^{N} A_i(z-1)^{N-i}}$$

The other methods may be attributed to the work of Jantzi, Snelgrove, Ferguson, Thurston, Pearce, Hawksford, and others [JSF93], [Ge89], [TPH91], who realized STF and NTFs with non-Butterworth responses using a chain of integrators with feedforward and feedback paths containing local resonator feedback loops. These methods have proven useful in implementing both bandpass signal and bandstop noise transfer functions and will be examined in the succeeding subsection on bandpass conversion design.

For analytical purposes, multiloop and cascaded sigma-delta converters may be accurately modelled as a linear subsystem with a nonlinear operator (the quantizer). Therefore, the quantizer may be represented as an additive error signal source provided that no prior assumptions are made regarding the behavior of this error signal source (*e.g.* assuming quantizer error behaviour as an independent identically distributed noise source). In this way the $\Sigma - \Delta$ converter may be characterized by the following input-to-output relationship in the z-domain

$$Y(z) = H_S(z)U(z) + H_N(z)E(z) \tag{1.3}$$

where $U(z)$ denotes the z-transformed input signal and $E(z)$ denotes the z-transformed quantization error signal, and where $Y(z)$ denotes the corresponding z-transformed converter output signal. Moreover, the signal transfer function is denoted as $H_S(z)$ and the noise transfer function by $H_N(z)$.

For the single-loop converter shown in Fig. 1.5, the converter may be analyzed by using Mason's gain formula to obtain the converter output response. Using the method outlined above, the output signal $Y(z)$ of the single-loop converter may be obtained as

$$Y(z) = G_1 z^{-1} U(z) + (1 - G_1 z^{-1}) E_1(z) \tag{1.4}$$

By setting the multiplier gain G_1 to unity (the ideal condition where multipliers would not be required), the equation reduces to

$$Y(z) = z^{-1} U(z) + (1 - z^{-1}) E_1(z) \tag{1.5}$$

It may be clearly observed from (1.5), that $H_S(z) = z^{-1}$ has an allpass nature, while $H_N(z) = (1 - z^{-1})$ has a highpass nature.

Other examples of chain of integrator structures are the basic second-order converter [Can85] shown in Fig. 1.9. This converter can be analysed to give

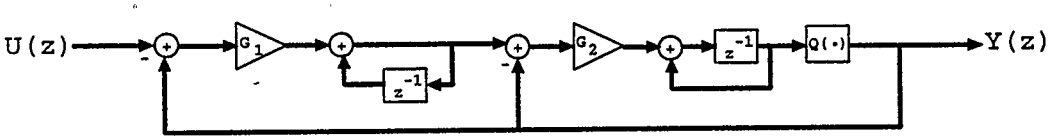


Figure 1.9. Double-loop $\Sigma - \Delta$ converter

$$\begin{aligned} Y(z) = & \frac{G_1 G_2 z^{-1}}{(1 - (1 - G_1 G_2 - G_2) z^{-1} + (1 - G_2) z^{-2})} U(z) \\ & + \frac{(1 - z^{-1})^2}{(1 - (1 - G_1 G_2 - G_2) z^{-1} + (1 - G_2) z^{-2})} E_1(z) \end{aligned} \tag{1.6}$$

which may be further simplified by assuming ideal matching *i.e.* multipliers $G_1 = G_2 = 1$, to the form

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^2 E_1(z) \tag{1.7}$$

The basic third-order converter in Fig. 1.10 can be similarly analysed to obtain

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^3 E_1(z) \tag{1.8}$$

This converter configuration provides the desired noise shaping function of $(1-z^{-1})^3$.

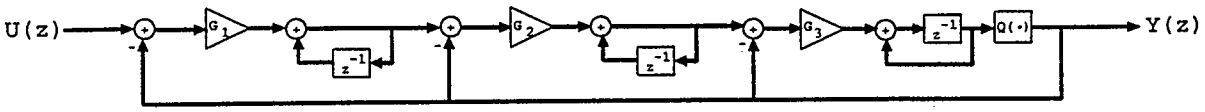


Figure 1.10. Triple-loop $\Sigma - \Delta$ converter

In the design of higher-order multiloop converters, only a few design methods besides that of cascade structures have been used successfully to implement cicuits with more than 3 loops. The design method by Chao, *et al.* is the most prevalent of these. There have been other methods employed that also utilize feedforward and feedback techniques, but they may be shown to be variations of the Chao, *et al.* design method only.

This section will examine the performance of the basic version of the Nth order loop topology sigma-delta converter proposed by Chao, *et al.*, the Triple Order Single Loop All Pole (TOSLAP) converter [CNLS90], where all the feedback coefficients (B_i's) are set to zero resulting in the simplified converter shown in Fig. 1.11. The converter configuration was analysed according to the original design specifications where: $A_0 = 0.8653$, $A_1 = 1.1920$, $A_2 = 0.3906$, $A_3 = 0.06926$, and $A_4 = 0.005395$. After performing a signal flow graph analysis of the TOSLAP converter to determine

the signal and noise transfer functions, and simplifying the result, the output equation for this converter configuration was found to be

$$Y(z) = \frac{0.8653z^{-1} - 2.2692z^{-2} + 2.0064z^{-3} - 0.59714z^{-4} + 0.000035z^{-5}}{1 - 3.1347z^{-1} + 3.7308z^{-2} - 1.9936z^{-3} + 0.40286z^{-4} + 0.000035z^{-5}} U(z)$$
$$+ \frac{1 - 4z^{-1} + 6z^{-2} - 4z^{-3} + z^{-4}}{1 - 3.1347z^{-1} + 3.7308z^{-2} - 1.9936z^{-3} + 0.40286z^{-4} + 0.000035z^{-5}} E_1(z) \quad (1.9)$$

It may be observed from this equation that the noise shaping performance of $H_N(z)$

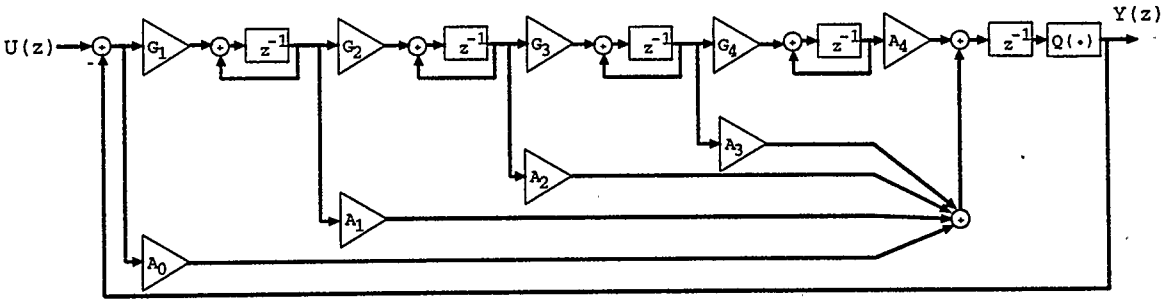


Figure 1.11. The triple order single loop all pole (TOSLAP) converter

in (1.9) approaches that of the desired fourth order NTF,

$$H_N(z) = (1 - z^{-1})^4$$

due to the placement of the NTF poles away from the origin.

1.3.2 Cascaded $\Sigma - \Delta$ converters

An alternative method for achieving high order noise shaping and therefore improved signal-to-quantization noise ratio and stability has been achieved with the use of cascaded low-order subsystems. It is well known that single-loop and double-loop $\Sigma - \Delta$ converters exhibit good stability properties [CT92], and with the cascade of such structures (with a modest increase in additional components), converters can be

constructed that achieve higher order noise shaping without the high parameter tolerances required for high-order multiloop converters. These converter configurations which contain multiple quantizers, are designed such that the quantizer error signal from the first through the $(N-1)$th stage (for an N stage converter configuration) are combined so their effects cancel one another. This results in only one source of quantizer error signal (the Nth) remaining at the output being shaped by the NTF

$$H_N(z) = (1 - z^{-1})^N. \tag{1.10}$$

In this manner, the noise shaping properties of higher-order multiloop converters may be obtained without their inherent stability problems due to high precision component requirements. A few converter configurations that have been designed using this technique will be presented next in order of increasing converter complexity, beginning with second-order converter configurations and ending with third-order converter configurations.

Second-order converter configurations were developed to exploit the second-order noise shaping of double-loop converters while retaining the stability properties of the basic first-order converter. In the field of converter design , there are different ideas as to how to achieve the desired output equation

$$Y(z) = z^{-N}U(z) + (1 - z^{-1})^2 E_2(z) \tag{1.11}$$

where N is the number of delays chosen by the designer for the overall output signal. There have been several different second-order converter configurations developed, such as those by Candy & Temes [CT92], or by Wong & Gray [WG90], and those by Uchimura, Hayashi, Kimura, and Iwata [UHKI88]. The signal and noise transfer functions introduced in the preceding section will be determined for each of these converter configurations in order to evaluate their performance characteristics.

The Candy & Temes converter is shown in Fig. 1.12. The overall output signal for this modulator assuming ideal matching ($G_1 = G_2 = 1$), is given by

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^2 E_2(z) \tag{1.12}$$

This converter performs the same as the double-loop converter of Fig. 1.9, while

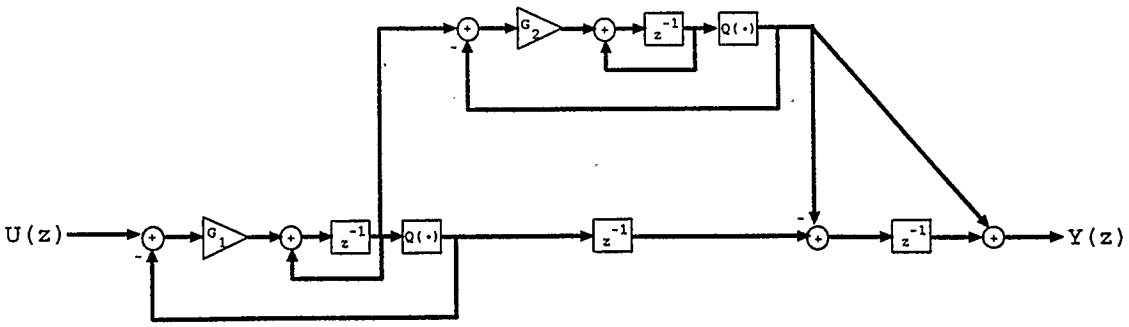


Figure 1.12. The cascade 2 converter by Candy & Temes

retaining the more stable structure of the basic first-order converter.

The Wong & Gray converter is shown in Fig. 1.13. By analysing the converter signal flow graph configuration, and assuming ideal matching ($G_1 = G_2 = 1$), one obtains the overall output signal as

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^2 E_2(z) \tag{1.13}$$

which is identical to that of the Candy & Temes converter in Fig. 1.12. The converter configuration proposed by Candy & Temes, however, requires one less adder for obtaining the second-order noise shaping transfer function.

The final cascade 2 converter to be examined was developed by Uchimura, *et al.* seen in Fig. 1.14. The overall output signal for this configuration assuming ideal matching is given by

$$Y(z) = z^{-2}(2 - z^{-1})U(z) + z^{-1}(1 - z^{-1})^2 E_1(z) + (1 - z^{-1})^2 E_2(z) \tag{1.14}$$

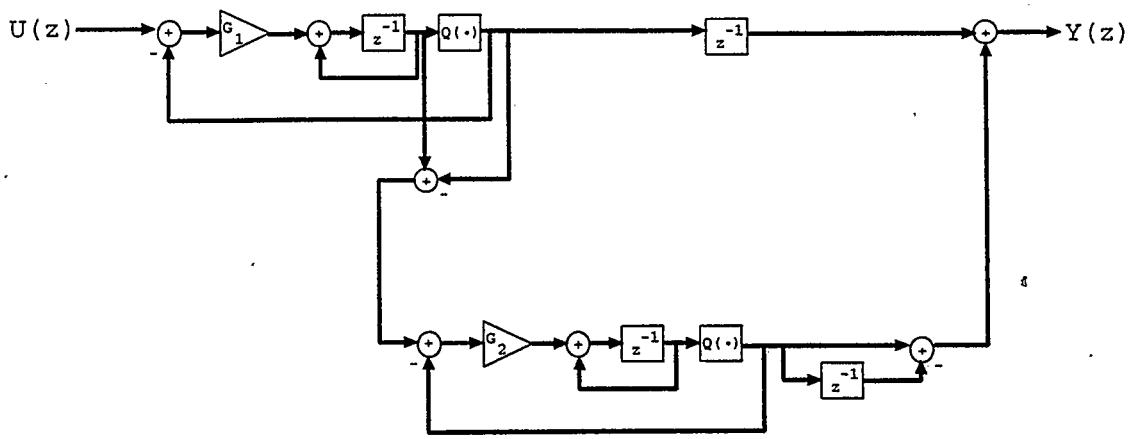


Figure 1.13. The cascade 2 converter by Wong & Gray

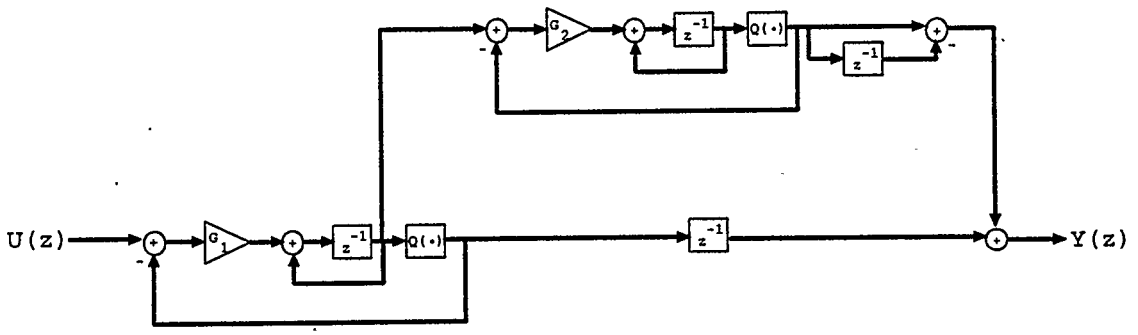


Figure 1.14. The cascade 2 converter by Uchimura, Hayashi, Kimura, and Iwata

This section will examine and compare the performance of cascade converter configurations of a more complex nature, that of third-order converter configurations. There have been several third-order cascade converters developed such as the Second Order First Order Cascade 1 (SOFOC1) converter [LC88], the Second Order First Order Cascade 2 (SOFOC2) converter [Rib91], and the Cascade21 converter [CT92]. The first of these converter configurations is the SOFOC1 seen in Fig. 1.15. The overall output signal for this configuration assuming ideal matching, is obtained as

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^3 E_2(z) \tag{1.15}$$

This converter provides the desired third-order noise shaping transfer function for a third-order converter. The next converter configuration to be examined is the SOFOC2 converter shown in Fig. 1.16. By analysing the converter signal flow graph configuration the overall output signal assuming ideal matching, is given by

$$\begin{aligned} Y(z) = &\frac{z^{-3}}{(1 - z^{-1} + z^{-2})} U(z) + \frac{z^{-2}(1 - z^{-1})^3}{(1 - z^{-1} + z^{-2})} E_1(z) \\ &+ \frac{(1 - 4z^{-1} + 7z^{-2} - 7z^{-3} + 4z^{-4} - z^{-5})}{(1 - z^{-1} + z^{-2})} E_2(z) \end{aligned} \tag{1.16}$$

This converter gives a decreased noise shaping performance (and therefore a reduced signal-to-quantization noise ratio) due to the placement of the poles of the noise transfer function away from the origin in (1.16). The final third-order converter to be examined is the Cascade21 converter configuration (see Fig. 1.17). This converter merely takes the output from the second integrator of a second-order converter (prior to quantization) and passes it through a second stage consisting of a first-order converter. The overall output signal for this converter assuming ideal matching, is given by

$$\begin{aligned} Y(z) = &\frac{2z^{-3} - 2z^{-4} + z^{-5}}{(1 - z^{-1} + z^{-2})} U(z) + \frac{z^{-1}(1 - z^{-1})^3}{(1 - z^{-1} + z^{-2})} E_1(z) \\ &+ \frac{(1 - 4z^{-1} + 7z^{-2} - 7z^{-3} + 4z^{-4} - z^{-5})}{(1 - z^{-1} + z^{-2})} E_2(z) \end{aligned} \tag{1.17}$$

It must be noted that the design of any cascaded $\Sigma - \Delta$ converter configurations must be done carefully as improper cancellation of quantization noise sources from prior stages preceding the final stage will result in converters with poor noise shaping

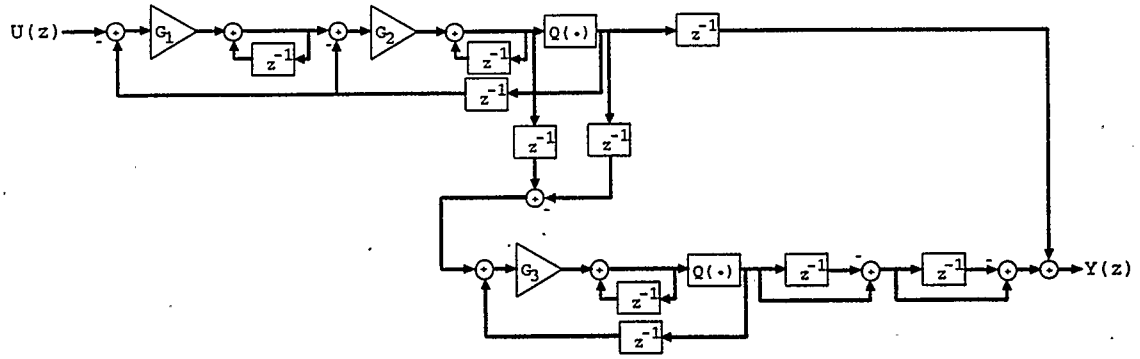


Figure 1.15. The second order first order cascade 1 (SOFOC1) converter

transfer functions or multiple sources of quantization error existing at the output, as seen in (1.14), (1.16) and (1.17).

1.4 Survey of design methods for bandpass $\Sigma - \Delta$ converters

Lowpass $\Sigma - \Delta$ converters have extremely low quantization noise only around DC [CB81]. By making use of this property, designers [SS89] have developed converters where quantization noise is reduced to zero at some frequency $\omega_0 = 2\pi f_0$ to obtain good noise suppression in a band around ω_0 (see Figs. 1.18a and 1.18b). Then, through the use of a narrow bandpass filter centered around ω_0 a bandpass $\Sigma - \Delta$ converter may be obtained for the suppression of noise around ω_0 and not DC.

One important difference between bandpass $\Sigma - \Delta$ conversion and that of lowpass $\Sigma - \Delta$ conversion is the definition of the oversampling ratio. The oversampling ratio for a lowpass $\Sigma - \Delta$ converter is defined as

$$OSR = \frac{f_s}{2f_0} \tag{1.18}$$

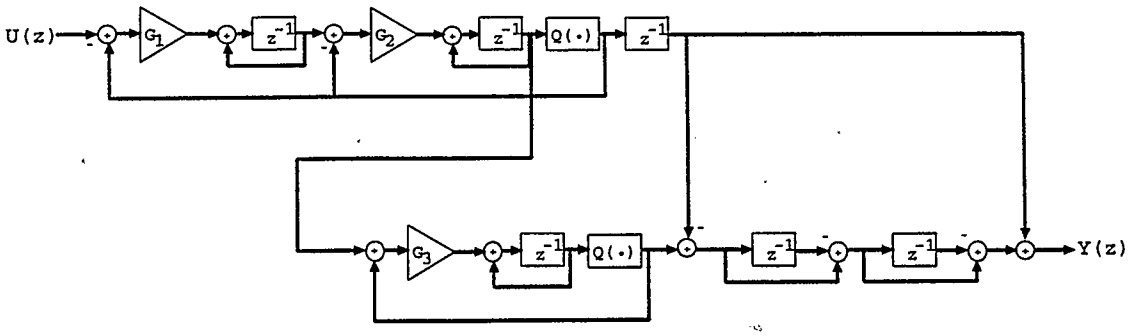


Figure 1.16. The second order first order cascade 2 (SOFOC2) converter

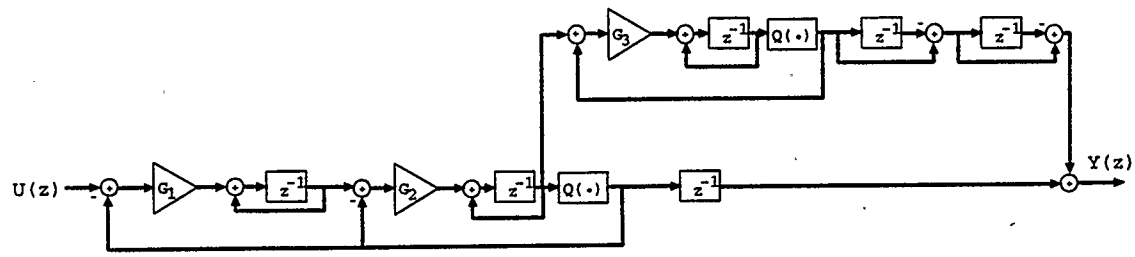
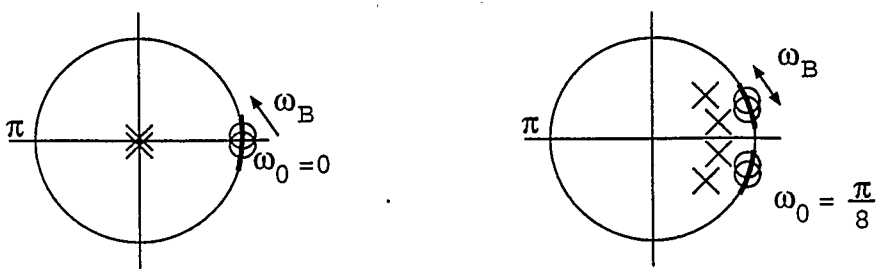


Figure 1.17. The cascade 21 converter



a) Lowpass Sigma-Delta Converter(Second order)

b) Bandpass Sigma-Delta Converter(Fourth order)

Figure 1.18. Comparison of pole and zero placements of NTF for lowpass and bandpass $\Sigma - \Delta$ converters (the passbands are highlighted)

while that for a bandpass $\Sigma - \Delta$ converter is given by

$$OSR = \frac{f_s}{2f_b} \tag{1.19}$$

where f_b is the frequency bandwidth of interest. Thus for a signal centered around 1 MHz with $f_b = 100kHz$ and $f_s = 5MHz$, the oversampling ratio would be 25 instead of 2.5 as in the case of a lowpass converter.

In bandpass $\Sigma - \Delta$ converters, the STF and NTF are chosen as follows. The STF must provide a unity gain in the desired passband (preferably a gain < 1 in frequencies outside the passband). The NTF is then chosen such that

a) large inband attenuation is obtained;

b) the out-of-band NTF gain is chosen to be less than 2 to keep the converter stable according to Lee's rule of thumb for 1 bit quantizers [CNLS90]; and

c) the first NTF impulse response coefficient is chosen to be unity to avoid delay-free loops and guarantee the realizability of the converter [JSS91].

By adopting these criteria together with other criteria such as trade-offs between sampling rate, oversampling ratio, and the anti-aliasing requirement for choosing

the location and width of the of the frequency band of interest, filter optimization routines [JSS91] are usually employed to obtain a circuit configuration suitable for implementation. To illustrate the point, the following bandpass $\Sigma - \Delta$ converter composed of a chain of integrators with feedforward and feedback coefficients and resonator feedback loops (see Fig. 1.19) was designed to produce the STF and NTF magnitude reponses shown in Fig. 1.20. These responses were useful in the design of an A/D converter for IF frequency applications [JSF93].

1.5 Overview of the thesis

In the years following the resurgence of research on $\Sigma - \Delta$ modulators as triggered in the mid 1970's by J. Candy [Can74] - [CB81], considerable effort has been made in examining trade-offs between the performance of $\Sigma - \Delta$ converters and their complexity. Numerous results have been obtained and reported through computer simulations, but these results only provide limited information regarding the actual behaviour of the considered $\Sigma - \Delta$ converters. It is more important, to develop analytical methods for the derivation of the relationships characterizing these converters, not only for a basic understanding of these converters but also for the development of novel type converters and their improvement.

The purpose of this thesis is twofold: (a) to introduce generalized state-space methods for the analysis and characterization of a class of $\Sigma - \Delta$ converters, and (b) to introduce several new $\Sigma - \Delta$ converters. To facilitate these objectives, this thesis is organized in the following manner.

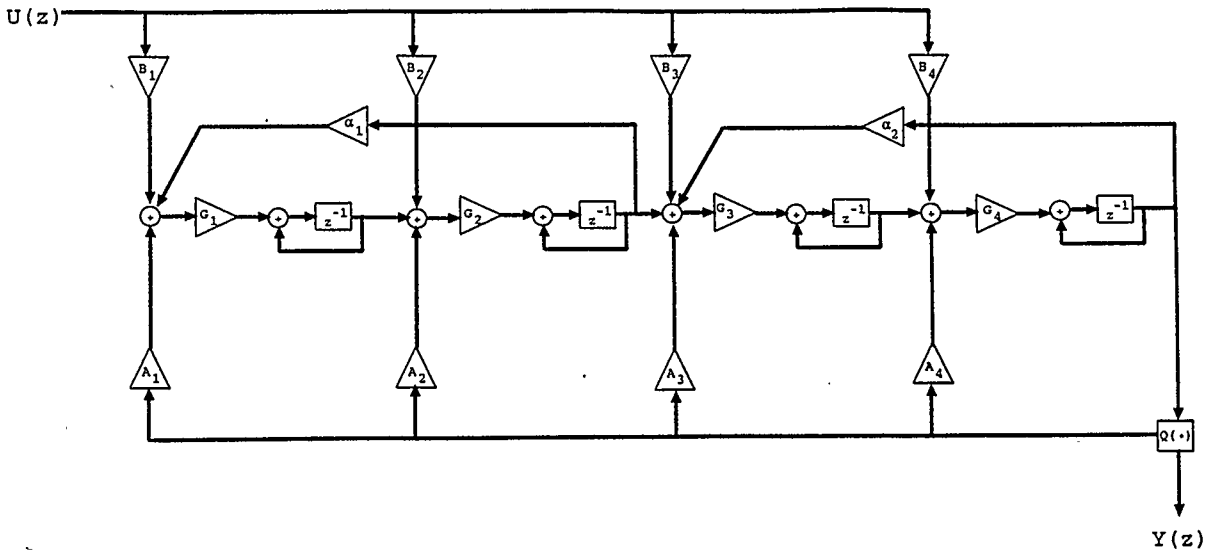


Figure 1.19. Fourth-order bandpass resonator converter

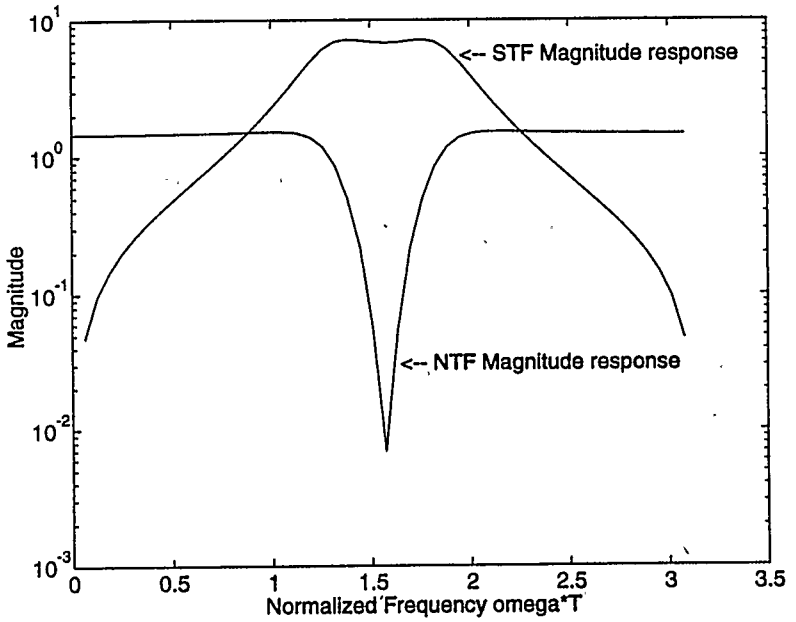


Figure 1.20. The STF and NTF magnitude/frequency responses for a fourth order bandpass converter

Chapter 2 will introduce a general state-space formulation for the representation of single-quantizer $\Sigma - \Delta$ converters. Using this formulation, a closed-form solution will be derived for the granular quantization error under the assumption of no-overload quantizer operation. An input signal bound will then be derived to guarantee the no-overload operation. The results are exploited to place in evidence interrelationships between converter parameters such as the input signal range, quantization bin-width, and number of quantization levels. Finally, several application examples are presented to illustrate the results.

Chapter 3 will utilize the proposed state-space representation to derive the closed-form solution for the quantizer output signal. This will then be followed by the derivation of an open-loop equivalent system. Proof by empirical results will then be presented using some application examples. Finally the stability aspects of $\Sigma - \Delta$ converters with regard to both circuit parameters and input signal range will be discussed together with several application examples.

Chapter 4 will examine $\Sigma - \Delta$ converter quantization noise spectra, beginning with a brief overview of several spectral analysis techniques, particularly the method for use with the linearized $\Sigma - \Delta$ converter model. Then more realistic methods taking into account the nonlinear nature of the quantizer error into account will be examined, resulting in a Fourier series expansion of the quantization error. Finally, this chapter will conclude with a calculation of the signal-to-quantization noise ratio for the three discussed methods.

Chapter 5 will discuss the desired characteristics of the signal and noise transfer functions in $\Sigma - \Delta$ converter design. Then several new $\Sigma - \Delta$ converters are presented and subsequently analysed with regard to signal and noise transfer functions. It will then characterize their performance with regard to signal-to-quantization noise ratio.

Chapter 6 will finally conclude the thesis with a brief overview of the material presented, a discussion of the fields of analysis and design of $\Sigma-\Delta$ converters that should be explored more fully in the future, and then present some concluding remarks.

CHAPTER 2

CLOSED-FORM SOLUTION OF GRANULAR QUANTIZATION ERROR FOR A CLASS OF SIGMA-DELTA CONVERTERS: A STATE-SPACE APPROACH

2.1 Introduction

The main objective of this chapter is to develop a closed-form solution for the granular quantization error to be subsequently used in the analysis of a widely used class of $\Sigma-\Delta$ converters. The development is facilitated by extracting the constituent quantizer from the $\Sigma - \Delta$ converter configuration, thus partitioning the converter configuration into a linear time-invariant subsystem and the quantizer itself. This partitioning leads in a straightforward manner to the derivation of a nonlinear matrix difference equation relating the quantizer error to its past values and the present and past values of the input signal via the arithmetic operation of the quantizer. The closed-form solution is subsequently obtained by solving this matrix equation for arbitrary input signals. This closed-from solution is derived under the assumption that the constituent quantizer operates in its no-overload region. To render the results of the closed-form solution complete, an input signal bound is derived to guarantee no-overload quantizer operation.

2.2 Formulation of the Problem Statement

A general $\Sigma - \Delta$ converter configuration can be represented as shown in Fig. 2.1. This representation is obtained by extracting the constituent quantizer $\mathcal{Q}$ from the $\Sigma - \Delta$ converter configuration, resulting in the identification of the linear time-invariant subsystem $\mathcal{N}$.

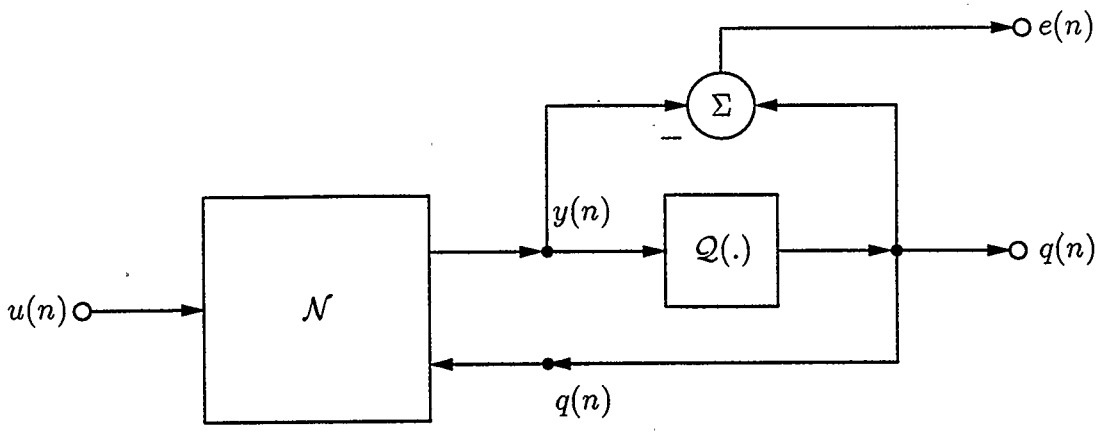


Figure 2.1. The general single quantizer converter representation

By using a state-space formulation for the linear time-invariant subsystem $\mathcal{N}$, the state and output equations for the subsystem $\mathcal{N}$ may be written as

$$x(n+1) = Ax(n) + B_1 u(n) + B_2 q(n) \tag{2.1}$$

$$y(n) = Cx(n) + Du(n) \tag{2.2}$$

where $x(n)$ represents the state vector and $u(n)$ represents the converter input signal, and where $y(n)$ represents the signal before and $q(n)$ represents the signal after quantization ($q(n)$ also represents the converter output signal). Moreover, A is an $N \times N$ matrix, B_1 and B_2 are $N \times 1$ vectors, C is a $1 \times N$ vector, with D being a scalar. Note that in order to render the overall $\Sigma - \Delta$ converter output $q(n)$ computable, it

has been implicitly assumed $y(n)$ is independent of the present value of $q(n)$ (cf. Fig. 2.1). Finally, the quantizer subsystem operates on its input according to the equation

$$q(y) = \begin{cases} (M-1)\Delta/2 & for \quad (M/2-1)\Delta \leq y \\ (k-1/2)\Delta & for \quad (k-1)\Delta \leq y < k\Delta \\ (-M+1)\Delta/2 & for \quad y < (-M/2+1)\Delta \end{cases} \tag{2.3}$$

where $k = (-M/2+2), \ldots, (M/2-1)$, where M represents the number of quantization levels, and where Δ represents the quantization bin-width [Gra90] (the separation between adjacent quantization levels). In this formulation, M is taken as an even number (the case of an odd M can be considered in the same manner).

In Fig. 2.1, the quantizer error signal is defined as

$$e(n) = q(n) - y(n) \tag{2.4}$$

If the quantizer input $y(n)$ is confined to the interval $[-M\Delta/2, M\Delta/2]$, then the quantizer error will be guaranteed to be bounded from above by $\Delta/2$, reducing to *granular* quantization error. Otherwise, the quantizer operates in the so-called overload region.

The problem under consideration is to derive a closed-form solution for the granular quantization error $e(n)$ by solving (2.1) and (2.2) in combination with (2.3).

2.3 Derivation of the Closed-form Solution of the Granular Quantization Error

The objective of this section is three-fold, a) to derive a nonlinear matrix equation characterizing the quantization error $e(n)$ for arbitrary input signals $u(n)$, b) to define the condition for no-overload quantizer operation, and c) to derive a closed-form solution for the granular quantization error $e(n)$ (i.e. the error associated with overload-free quantizer operation).

2.3.1 Nonlinear matrix difference equation

To begin with, (2.4) and (2.2) are used in (2.1) to obtain

$$x(n+1) = (A + B_2C)x(n) + (B_1 + B_2D)u(n) + B_2e(n) \tag{2.5}$$

By solving this equation recursively and by substituting the result in (2.2), one obtains

$$\begin{aligned} y(n) = {} & C(A + B_2C)^n x(0) + C\sum_{i=0}^{n-1}(A + B_2C)^i(B_1 + B_2D)u(n-i-1) \\ & + C\sum_{i=0}^{n-1}(A + B_2C)^i B_2 e(n-i-1) + Du(n) \end{aligned} \tag{2.6}$$

Furthermore, the quantizer operation is represented by

$$\begin{aligned} q(n) &= q(\frac{1}{\Delta}y(n)) \\ &= \frac{\Delta}{2} + \Delta\lfloor\frac{1}{\Delta}y(n)\rfloor \end{aligned} \tag{2.7}$$

where $\lfloor\cdot\rfloor$ denotes the floor of its argument. Similarly, $y(n)$ is represented by

$$y(n) = \Delta\lfloor\frac{1}{\Delta}y(n)\rfloor + \Delta\langle\frac{1}{\Delta}y(n)\rangle \tag{2.8}$$

where $\langle\cdot\rangle$ represents the fractional part of its argument. Then, by substituting (2.7) and (2.8) into (2.4), one obtains

$$e(y) = \frac{\Delta}{2} - \Delta\langle\frac{1}{\Delta}y(n)\rangle \tag{2.9}$$

Finally, the substitution of (2.6) into (2.9), yields

$$\begin{aligned} e(n) = \frac{\Delta}{2} - \Delta\Bigg\langle\frac{1}{\Delta}\Bigg[& C(A + B_2C)^n x(0) \\ & + C\sum_{i=0}^{n-1}(A + B_2C)^i(B_1 + B_2D)u(n-i-1) \\ & + C\sum_{i=0}^{n-1}(A + B_2C)^i B_2 e(n-i-1) + Du(n)\Bigg]\Bigg\rangle \end{aligned} \tag{2.10}$$

2.3.2 Closed-form solution

Definition 1 *If the input signal $y(n)$ to the quantizer $\mathcal{Q}$ in the $\Sigma - \Delta$ modulator in Fig. 1 is such that*

$$-M\Delta/2 \leq y(n) \leq M\Delta/2 \tag{2.11}$$

for $n = 0, 1, 2, \ldots, \infty$, then $\mathcal{Q}$ is said to operate in its no-overload region.

In the no-overload region, the quantization error $e(n)$ becomes granular in accordance with

$$-\Delta/2 \leq e(n) \leq \Delta/2$$

for $n = 0, 1, 2, \ldots, \infty$.

The closed-form solution for the granular quantization error is given in the following theorem.

Theorem 2.1 ([BN95]) *If the scalars $C_{1\times N}A^l_{N\times N}B_{2N\times 1}$ are integral numbers for each $l = 0, 1, \ldots, \infty$, then the closed-form solution of (2.10) is given by*

$$\begin{aligned} e(n) = \frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\left[CA^n x(0) + C\sum_{i=0}^{n-1} A^i B_1 u(n-i-1)\right.\right. \\ \left.\left. + C\sum_{i=0}^{n-1} A^i B_2 \frac{\Delta}{2} + Du(n)\right]\right\rangle \end{aligned} \tag{2.12}$$

The proof of Theorem 2.1 will be given after establishing Lemmas 1, 2, and 3 below.

Lemma 1 *For any integer $n \geq 1$,*

$$(A + B_2C)^n - \sum_{i=0}^{n-1}(A + B_2C)^i B_2 C A^{n-i-1} = A^n \tag{2.13}$$

Proof: By induction. The lemma is clearly valid for $n = 1$. Therefore, it is sufficient to show that if it holds true for $n = m$, then it is also valid for $n = m + 1$.

For $n = m + 1$, the left-hand side of (2.13) can be written as

$$\begin{aligned}(A + B_2C)^{m+1} - \sum_{i=0}^{m}(A + B_2C)^i B_2C A^{m-i} \\ = (A + B_2C)^m A - \left[\sum_{i=0}^{m-1}(A + B_2C)^i B_2C A^{m-i-1}\right] A\end{aligned} \tag{2.14}$$

But, by setting $n = m$ in (2.13), one has

$$(A + B_2C)^m - \sum_{i=0}^{m-1}(A + B_2C)^i B_2C A^{m-i-1} = A^m \tag{2.15}$$

Then, by substituting (2.15) in (2.14), and by simplifying the result, one arrives at

$$(A + B_2C)^{m+1} - \sum_{i=0}^{m}(A + B_2C)^i B_2C A^{m-i} = A^{m+1} \tag{2.16}$$

which shows that the lemma also holds true for $n = m + 1$. □

Lemma 2 *For any integer* $n \geq 0$,

$$\sum_{i=0}^{n}(A + B_2C)^i - \sum_{i=0}^{n}(A + B_2C)^i B_2C \sum_{k=0}^{n-i-1} A^k = \sum_{i=0}^{n} A^i \tag{2.17}$$

Proof: By induction. The lemma clearly holds true for $n = 0$. Therefore, it suffices to show that if it is valid for $n = m$, then it also holds true for $n = m + 1$.

For $n = m + 1$, the left-hand side of (2.17) can be written in the form

$$\begin{aligned}&\sum_{i=0}^{m+1}(A + B_2C)^i - \sum_{i=0}^{m+1}(A + B_2C)^i B_2C \sum_{k=0}^{m-i} A^k \\ &= (A + B_2C)^{m+1} + \sum_{i=0}^{m}(A + B_2C)^i \\ &\quad - \sum_{i=0}^{m+1}(A+B_2C)^i B_2C \sum_{k=0}^{m-i-1} A^k - \sum_{i=0}^{m+1}(A+B_2C)^i B_2C A^{m-i}\end{aligned} \tag{2.18}$$

But, by setting $n = m$ in (2.17), one gets

$$\sum_{i=0}^{m}(A + B_2C)^i - \sum_{i=0}^{m}(A + B_2C)^i B_2C \sum_{k=0}^{m-i-1} A^k = \sum_{i=0}^{m} A^i \tag{2.19}$$

Then, by making use of (2.19) and Lemma 1 for $n = m$, the right-hand side of (2.18) simplifies to

$$\sum_{i=0}^{m} A^i + A^{m+1} = \sum_{i=0}^{m+1} A^i \tag{2.20}$$

which shows that the lemma is also valid for $n = m + 1$. □

Lemma 3 *For any integer* $n \geq 0$,

$$\begin{aligned} &\sum_{i=0}^{n}(A + B_2C)^i B_1 u(n-i) \\ &\quad - \sum_{i=0}^{n}(A + B_2C)^i B_2C \sum_{k=0}^{n-i-1} A^k B_1 u(n-i-k-1) \\ &= \sum_{i=0}^{n} A^i B_1 u(n-i) \end{aligned} \tag{2.21}$$

Proof: By induction. The lemma is clearly valid for $n = 0$. Therefore, it is sufficient to show that if it holds true for $n = m$, then it is also valid for $n = m + 1$.

For $n = m + 1$, the left-hand side of (2.21) can be manipulated as

$$\begin{aligned} &\sum_{i=0}^{m+1}(A + B_2C)^i B_1 u(m-i+1) \\ &\quad - \sum_{i=0}^{m+1}(A + B_2C)^i B_2C \sum_{k=0}^{m-i} A^k B_1 u(m-i-k) \\ &= \sum_{i=0}^{m+1}(A + B_2C)^i B_1 u(m-i+1) \\ &\quad - \sum_{i=0}^{m+1}\sum_{k=0}^{i}(A + B_2C)^k B_2C A^{i-k} B_1 u(m-i+1) \end{aligned} \tag{2.22}$$

Then, by invoking Lemma 1, the right-hand side of (2.22) simplifies to

$$\begin{aligned} &\sum_{i=0}^{m+1}\left[(A + B_2C)^i - \sum_{k=0}^{i}(A + B_2C)^k B_2C A^{i-k}\right] B_1 u(m+1-i) \\ &\quad = \sum_{i=0}^{m+1} A^i B_1 u(m+1-i) \end{aligned} \tag{2.23}$$

which shows that the lemma also holds true for $n = m + 1$. □

Proof of Theorem 2.1. By induction. The theorem clearly holds true for $n = 0$. In this way, it is sufficient to show that if is valid for $n = 0, 1, 2, \ldots, m$, then it also holds true for $n = m + 1$.

By setting $n = m + 1$ in (2.10), one has

$$\begin{aligned} e(m+1) = \frac{\Delta}{2} - \Delta \left\langle \frac{1}{\Delta} \Big[C(A+B_2C)^{m+1}x(0) \right. \\ + C\sum_{i=0}^{m}(A+B_2C)^i(B_1+B_2D)u(m-i) \\ \left. + C\sum_{i=0}^{m}(A+B_2C)^iB_2e(m-i) + Du(m+1) \Big] \right\rangle \end{aligned} \tag{2.24}$$

By invoking (2.12) for the terms $e(m-i)$ in (2.24), and utilizing the property

$$\langle r_1 + K\langle r_2 \rangle \rangle = \langle r_1 + Kr_2 \rangle \tag{2.25}$$

for real numbers r_1 and r_2 and integer number K in the result, one can obtain

$$\begin{aligned} e(m+1) = \frac{\Delta}{2} - \Delta \left\langle \frac{1}{\Delta} \Big[C(A+B_2C)^{m+1}x(0) \right. \\ + C\sum_{i=0}^{m}(A+B_2C)^i(B_1+B_2D)u(m-i) \\ + C\sum_{i=0}^{m}(A+B_2C)^iB_2\frac{\Delta}{2} - C\sum_{i=0}^{m}(A+B_2C)^iB_2CA^{m-i}x(0) \\ - C\sum_{i=0}^{m}(A+B_2C)^iB_2C\sum_{k=0}^{m-i-1}A^kB_1u(m-i-k-1) \\ - C\sum_{i=0}^{m}(A+B_2C)^iB_2C\sum_{k=0}^{m-i-1}A^kB_2\frac{\Delta}{2} \\ \left. - C\sum_{i=0}^{m}(A+B_2C)^iB_2Du(m-i) + Du(m+1) \Big] \right\rangle \end{aligned} \tag{2.26}$$

By invoking Lemmas 1, 2, and 3 above in (2.26), and some tedious manipulation, one arrives at

$$\begin{aligned} e(m+1) = \frac{\Delta}{2} - \Delta \left\langle \frac{1}{\Delta} \Big[CA^{m+1}x(0) + C\sum_{i=0}^{m}A^iB_1u(m-i) \right. \\ \left. + C\sum_{i=0}^{m}A^iB_2\frac{\Delta}{2} + Du(m+1) \Big] \right\rangle \end{aligned} \tag{2.27}$$

which shows that the theorem is also valid for $n = m + 1$. □

Using the necessary condition that the matrix product must be an integer for all i, we may define the class of $\Sigma - \Delta$ converters for which Theorem 2.1 holds as being those single-quantizer converters that contain integer valued multipliers. This class therefore includes the single, double, and triple-loop converters as special cases.

2.4 Derivation of the input signal bound for overload-free $\Sigma - \Delta$ converter operation

The closed-form solution in (2.12) for the $\Sigma - \Delta$ converter *granular* quantization error was derived under the assumption that the constituent quantizer operates in the no-overload region. The objective of this section is to derive a bound on the $\Sigma - \Delta$ converter input signal $u(n)$ for the required overload-free quantizer operation. The result is first obtained for the case of zero initial state $x(0)$, followed by the case of nonzero initial state $x(0)$.

In accordance with (2.6), the quantizer input may be represented by

$$y(n) = C(A + B_2C)^n x(0) + C\sum_{i=0}^{n-1}(A + B_2C)^i(B_1 + B_2D)u(n-i-1)$$
$$+ C\sum_{i=0}^{n-1}(A + B_2C)^i B_2 e(n-i-1) + Du(n) \tag{2.28}$$

Let the matrix $(A + B_2C)$ be stable, *i.e.* let the magnitudes of its eigenvalues be strictly less than unity. Then it can be shown that for a bounded initial state vector, a bounded input u_i (for $n = 0, 1, 2, \ldots, \infty$), and a bounded error e_i (for $n = 0, 1, 2, \ldots, \infty$), one produces a bounded output $y(n)$ from the linear time-invariant subsystem $\mathcal{N}$ which is also the quantizer input. Now by using the property

$$\|AB\| \leq \|A\| \cdot \|B\|$$

for a matrix A and a vector B (having compatible norms [HJ93]) and by using the

Triangle Inequality, (2.28) may be rewritten as

$$\begin{aligned}\|y(n)\| &\leq \|C\| \cdot \|(A+B_2C)^n x(0)\| + \|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i (B_1+B_2D)u(n-i-1)\| \\ &\quad + \|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i B_2 e(n-i-1)\| + \|Du(n)\| \\ &\leq \|C\| \cdot \|(A+B_2C)^n\| \cdot \|x(0)\| \\ &\quad + \|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i\| \cdot \|(B_1+B_2D)\| \cdot \|u(n-i-1)\| \\ &\quad + \|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i\| \cdot \|B_2\| \cdot \|e(n-i-1)\| + \|D\| \cdot \|u(n)\|\end{aligned} \tag{2.29}$$

If the input sequence is bounded, there exists a finite number U such that $\|u_i\| < U$ for $i = 0, 1, 2, \ldots$, and if the error sequence is bounded, then $\|e_i\| < \Delta/2$ for $i = 0, 1, 2, \ldots$. Consequently, (2.29) may be rewritten as

$$\begin{aligned}\|y(n)\| &\leq \|C\| \cdot \|(A+B_2C)^n\| \cdot \|x(0)\| + U\|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i\| \cdot \|(B_1+B_2D)\| \\ &\quad + \frac{\Delta}{2}\|C\| \sum_{i=0}^{n-1} \|(A+B_2C)^i\| \cdot \|B_2\| + U\|D\|\end{aligned} \tag{2.30}$$

As the matrix $(A+B_2C)$ is stable, it can be shown that there exists a finite number W, and a positive real number ρ such that [ZD63]

$$\|(A+B_2C)^i\| < W\rho^i \qquad i = 0, 1, 2, \ldots \tag{2.31}$$

Moreover, $\rho < 1$ but is greater than the magnitudes of all the eigenvalues of $(A+B_2C)$. Consequently

$$\sum_{i=0}^{n} \|(A+B_2C)^i\| < W(1-\rho^{n+1})(1-\rho)^{-1} < W(1-\rho)^{-1} \tag{2.32}$$

In this way, the first term in the right-hand side of (2.30) is bounded by $W\|C\| \cdot \|x(0)\|$ for all n, the second by $U\|(B_1+B_2D)\|W(1-\rho)^{-1}$, the third by $(\Delta/2)\|B_2\|W(1-\rho)^{-1}$ and the final term by $U\|D\|$. Therefore,

$$\begin{aligned}\|y(n)\| &\leq W\|C\| \cdot \|x(0)\| + U\|(B_1+B_2D)\|W(1-\rho)^{-1} \\ &\quad + \frac{\Delta}{2}\|B_2\|W(1-\rho)^{-1} + U\|D\|\end{aligned} \tag{2.33}$$

2.4.1 For the case: $x(0) = 0$

For the case where the initial states are set to zero, the quantizer input signal may be shown to be

$$\begin{aligned} y(n) = C \sum_{i=0}^{n-1} (A + B_2C)^i (B_1 + B_2D) u(n-i-1) \\ + C \sum_{i=0}^{n-1} (A + B_2C)^i B_2 e(n-i-1) + Du(n) \end{aligned} \tag{2.34}$$

By using this equation, the input signal bound for the no-overload condition is obtained as given in the following theorem.

Theorem 2.2 *If $(A+B_2C)$ is a stable matrix, and if the input signal $u(n)$ is confined to the range $u(n) \in [-U_1, U_1]$ with*

$$U_1 \leq \frac{(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \frac{\Delta}{2} \tag{2.35}$$

then the quantizer $\mathcal{Q}$ will operate in its no-overload region for $n = 0, 1, 2, \ldots, \infty$, where W and ρ are as stated previously.

The proof of Theorem 2 will be given after establishing the following lemma.

Lemma 4

$$\frac{\|D\|(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \frac{\Delta}{2} \leq \frac{M\Delta}{2} \tag{2.36}$$

Proof: By contradiction. Let

$$\frac{\|D\|(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \frac{\Delta}{2} > \frac{M\Delta}{2} \tag{2.37}$$

Then,

$$-\|D\| \cdot \|C\| \cdot \|B_2\| W(1-\rho)^{-1} \frac{\Delta}{2} > \|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} \frac{M\Delta}{2} \tag{2.38}$$

which is a contradiction, as a negative number cannot be greater than a positive number. □

Proof of Theorem 2.2: By induction. For $n = 0$, from (2.32) and (2.34)

$$\|y(0)\| = \|D\| \cdot \|u(0)\| \leq \|D\|U_1 \leq \frac{\|D\|(M - \|C\| \cdot \|B_2\|W(1-\rho)^{-1})}{(\|C\| \cdot \|(B_1 + B_2D)\|W(1-\rho)^{-1} + \|D\|)} \frac{\Delta}{2} \tag{2.39}$$

By making use of Lemma 4, this reduces to $\|y(0)\| \leq M\Delta/2$ yielding $\|e(0)\| \leq \Delta/2$. Therefore, it is sufficient to show that if $-\Delta/2 \leq e(n) \leq \Delta/2$ is true for $n = 0, 1, 2, \ldots, m-1$, then $-\Delta/2 \leq e(n) \leq \Delta/2$ is also valid for $n = m$.

For $n = m$, from (2.34) one gets

$$\begin{aligned} \|y(m)\| \leq \|C\| \sum_{i=0}^{m-1} \|(A + B_2C)^i\| \cdot \|(B_1 + B_2D)\|U_1 \\ + \|C\| \sum_{i=0}^{m-1} \|(A + B_2C)^i\| \cdot \|B_2\| \frac{\Delta}{2} + \|D\|U_1 \leq \frac{M\Delta}{2} \end{aligned} \tag{2.40}$$

By using (2.32) in (2.40), one obtains

$$\begin{aligned} \|y(m)\| \leq \|C\|W(1-\rho)^{-1}\|(B_1 + B_2D)\|U_1 \\ + \|C\|W(1-\rho)^{-1}\|B_2\| \frac{\Delta}{2} + \|D\|U_1 \leq \frac{M\Delta}{2} \end{aligned} \tag{2.41}$$

Moreover, by invoking (2.35) in (2.41) one obtains $\|y(m)\| \leq M\Delta/2$, yielding $-\Delta/2 \leq e(m) \leq \Delta/2$. □

2.4.2 For the case: $x(0) \neq 0$

In accordance with (2.6), the quantizer input may be represented by

$$\begin{aligned} y(n) = C(A + B_2C)^n x(0) + C \sum_{i=0}^{n-1} (A + B_2C)^i (B_1 + B_2D)u(n-i-1) \\ + C \sum_{i=0}^{n-1} (A + B_2C)^i B_2 e(n-i-1) + Du(n) \end{aligned} \tag{2.42}$$

By using this equation, the input signal bound for the no-overload condition is obtained as given in the following theorem.

Theorem 2.3 *If $(A+B_2C)$ is a stable matrix, and if the input signal $u(n)$ is confined to the range $u(n) \in [-U_2, U_2]$ with*

$$U_2 \leq \frac{(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})\frac{\Delta}{2} - G}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \tag{2.43}$$

then the quantizer $\mathcal{Q}$ will operate in its no-overload region for $n = 0, 1, 2, \ldots, \infty$, where W and ρ are as stated previously, and where

$$G = W\|C\| \cdot \|x(0)\| \tag{2.44}$$

The proof of Theorem 2 will be given after establishing the following lemma.

Lemma 5

$$\frac{\|D\|[(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})\frac{\Delta}{2} - G]}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \leq \frac{M\Delta}{2} \tag{2.45}$$

Proof: By contradiction. Let

$$\frac{\|D\|[(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})\frac{\Delta}{2} - G]}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} > \frac{M\Delta}{2} \tag{2.46}$$

Then,

$$\begin{aligned} &- \|D\| \cdot \|C\| \cdot \|B_2\| W(1-\rho)^{-1}\frac{\Delta}{2} - \|D\|G \\ &\quad > \|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1}\frac{M\Delta}{2} \end{aligned} \tag{2.47}$$

which is a contradiction, as a negative number cannot be greater than a positive number. □

Proof of Theorem 2.3: By induction. For $n = 0$, from (2.42) and (2.32)

$$\begin{aligned} \|y(0)\| &= \|D\| \cdot \|u(0)\| \leq \|D\| U_2 \\ &\leq \frac{\|D\|[M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1}]\frac{\Delta}{2} - G}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)} \end{aligned} \tag{2.48}$$

By making use Lemma 5, this reduces to $\|y(0)\| \leq M\Delta/2$ yielding $\|e(0)\| \leq \Delta/2$. Therefore, it is sufficient to show that if $-\Delta/2 \leq e(n) \leq \Delta/2$ is true for $n = 0, 1, 2, \ldots, m-1$, then $-\Delta/2 \leq e(n) \leq \Delta/2$ is also valid for $n = m$.

For $n = m$, from (2.42) one gets

$$\begin{aligned} \|y(m)\| &\leq \|C\| \cdot \|(A + B_2C)^m\| \cdot \|x(0)\| \\ &+ \|C\| \sum_{i=0}^{m-1} \|(A + B_2C)^i\| \cdot \|(B_1 + B_2D)\|U_2 \\ &+ \|C\| \sum_{i=0}^{m-1} \|(A + B_2C)^i\| \cdot \|B_2\|\frac{\Delta}{2} + \|D\|U_2 \leq \frac{M\Delta}{2} \end{aligned} \tag{2.49}$$

By using (2.32) and (2.44) in (2.49), one obtains

$$\begin{aligned} \|y(m)\| &\leq G + \|C\|W(1-\rho)^{-1}\|(B_1 + B_2D)\|U_2 \\ &+ \|C\|W(1-\rho)^{-1}\|B_2\|\frac{\Delta}{2} + \|D\|U_2 \leq \frac{M\Delta}{2} \end{aligned} \tag{2.50}$$

Moreover, by invoking (2.43) in (2.50) one obtains $\|y(m)\| \leq M\Delta/2$, yielding $-\Delta/2 \leq e(m) \leq \Delta/2$. □

2.5 Interrelationships between the input signal bound and quantizer parameters

By using the above bounds and by using the facts that $M = 2^w$ and $\Delta = 2^{-(w-1)}$, some interesting relationships can be obtained for $\Sigma - \Delta$ converters. From (2.35), it can be shown that the minimum required number of quantizer levels must satisfy the relationship

$$\frac{M - \|C\| \cdot \|B_2\|W(1-\rho)^{-1}}{\|C\| \cdot \|(B_1 + B_2D)\|W(1-\rho)^{-1} + \|D\|} > 0 \tag{2.51}$$

or equivalently the minimum required wordlength w must satisfy the relationship.

$$\frac{2^w - \|C\| \cdot \|B_2\|W(1-\rho)^{-1}}{\|C\| \cdot \|(B_1 + B_2D)\|W(1-\rho)^{-1} + \|D\|} > 0 \tag{2.52}$$

Moreover, in design situations where the $\Sigma - \Delta$ converter is preceded by a saturator to limit the dynamic input signal range, the clipping levels may be chosen quite easily

in accordance with

$$\pm V_{max} = \pm \frac{(2^w - \|C\| \cdot \|B_2\| W(1-\rho)^{-1})}{(\|C\| \cdot \|(B_1 + B_2 D)\| W(1-\rho)^{-1} + \|D\|)} \frac{2^{-(w-1)}}{2} \tag{2.53}$$

if condition (2.52) is met.

2.6 Application Examples

In this section, the above results are applied to the determination of the input signal bound U_1 and the closed-form solution of the quantization error for conventional single, double, and triple-loop $\Sigma - \Delta$ converters given AC input signals having the form

$$u(n) = \alpha \cos(n 2\pi f / f_s + \theta) = \alpha \cos(n\omega + \theta)$$

and DC input signals having the form

$$u(n) = X, \text{ for all } n = 0, 1, 2, \ldots$$

To simplify matters, it is assumed that $x(0) = 0$ throughout.

2.6.1 Single-loop $\Sigma - \Delta$ converter

For the single-loop converter in Fig. 1.5, the subsystem $\mathcal{N}$ is represented by the state equations.

$$x(n+1) = [1]x(n) + [1]u(n) + [-1]q(n)$$

$$y(n) = [1]x(n) + (0)u(n)$$

By comparing these equations with the general form in (2.1) and (2.2), to identify the matrix A, and the vectors B_1, B_2, C, and D, and by using the results in (2.35),

one obtains the general input signal bound

$$U_1 \leq (M-1)\frac{\Delta}{2}$$

By choosing the number of quantizer levels as $M = 2$ together with a bin-width of $\Delta = 1$, one obtains the exact input signal bound $U_1 \leq 1/2$ and the exact quantizer no-overload region as $\|y(n)\| \leq 1$ from (2.33).

Verification of overload-free quantizer operation, given the above choices for M and Δ, may be seen in Fig. 2.2 for AC and DC inputs. By substituting for the matrix A, and vectors B_1, B_2, C, and D, the closed-form solution of the quantization error may be determined as

$$e(n) = \frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\left(\sum_{i=0}^{n-1}\left[u(n-i-1) - \frac{\Delta}{2}\right]\right)\right\rangle$$

Evaluating the above equation for an AC input signal and choosing $\Delta = 1$, results in the quantization error shown in Fig. 2.3b which matches exactly the quantization error obtained by using difference equations in Fig. 2.3a.

2.6.2 Double-loop $\Sigma - \Delta$ converter

For the double-loop converter of Fig. 1.9, the subsystem $\mathcal{N}$ is represented by the state equations

$$x(n+1) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} x(n) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(n) + \begin{bmatrix} -1 \\ -2 \end{bmatrix} q(n)$$

$$y(n) = \begin{bmatrix} 0 & 1 \end{bmatrix} x(n) + (0)u(n)$$

By comparing these equations with the general form in (2.1) and (2.2), to identify the matrix A, and the vectors B_1, B_2, C, and D, and by using these results in (2.35),

one obtains the general input signal bound

$$U_1 \leq (M-3)\frac{\Delta}{2}$$

Choosing the number of quantizer levels as $M = 4$ together with a bin-width of $\Delta = 1$, one obtains the exact input signal bound $U_1 \leq 1/2$ and the exact quantizer no-overload region as $\|y(n)\| \leq 2$ from (2.33).

Verification of overload-free quantizer operation, given the above choices for M and Δ, may be seen in Fig. 2.4 for AC and DC inputs. By substituting for the matrix A, and the vectors B_1, B_2, C, and D, the closed-form solution of the quantization error may be determined as

$$e(n) = \frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\left(\sum_{i=0}^{n-1}\left[(i+1)u(n-i-1) - \frac{(i+2)\Delta}{2}\right]\right)\right\rangle$$

Evaluating the above equation for an AC input signal and choosing $\Delta = 1$, results in the quantization error seen in Fig 2.5b which matches exactly the quantization error obtained by using difference equations in Fig. 2.5a.

2.6.3 Triple-loop $\Sigma - \Delta$ converter

For the triple-loop converter of Fig. 1.10, the subsystem $\mathcal{N}$ is represented by the state equations

$$x(n+1) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} x(n) + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} u(n) + \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix} q(n)$$

$$y(n) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x(n) + (0)u(n)$$

By comparing these equations with the general form in (2.1) and (2.2), to identify the matrix A, and the vectors B_1, B_2, C, and D, and by using these results in (2.35),

one obtains the general input signal bound

$$U_1 \leq (M-7)\frac{\Delta}{2}$$

Choosing the number of quantizer levels as $M = 8$ together with a bin-width of $\Delta = 1$, one obtains the exact input signal bound $U_1 \leq 1/2$ and the exact quantizer no-overload region as $\|y(n)\| \leq 4$ from (2.33).

Verification of overload-free quantizer operation, given the above choices for M and Δ, may be seen in Fig. 2.6 for AC and DC inputs. By substituting for the matrix A, and the vectors B_1, B_2, C, and D, the closed-form solution of the quantization error may be determined as

$$e(n) = \frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\left(\sum_{i=0}^{n-1}\left[\frac{(i+1)(i+2)}{2}u(n-i-1) - \frac{(i+2)(i+3)\Delta}{4}\right]\right)\right\rangle$$

Evaluating the above equation for an AC input signal and choosing $\Delta = 1$, results in the quantization error seen in Fig 2.7b which matches exactly the quantization error obtained by using difference equations in Fig. 2.7a.

2.7 Conclusion

This chapter has presented a closed-form solution for granular quantization error of a class of $\Sigma - \Delta$ converters using a state-space approach. This was accomplished by first developing a state-space representation of a single quantizer $\Sigma - \Delta$ converter. Then, the state-space equations describing this subsystem were used to derive a closed-form equation for the *granular* quantizer error under the no-overload condition. An input signal bound was then derived to guarantee no overloading of the quantizer for the two cases of a) $x(0) \neq 0$, and b) $x(0) = 0$. Some interrelationships between the input signal bound and quantizer parameters were then discussed, and finally the utility of the results obtained for the determination of the input signal

bound and the closed-form solution for the quantization error were illustrated through some practical application examples.

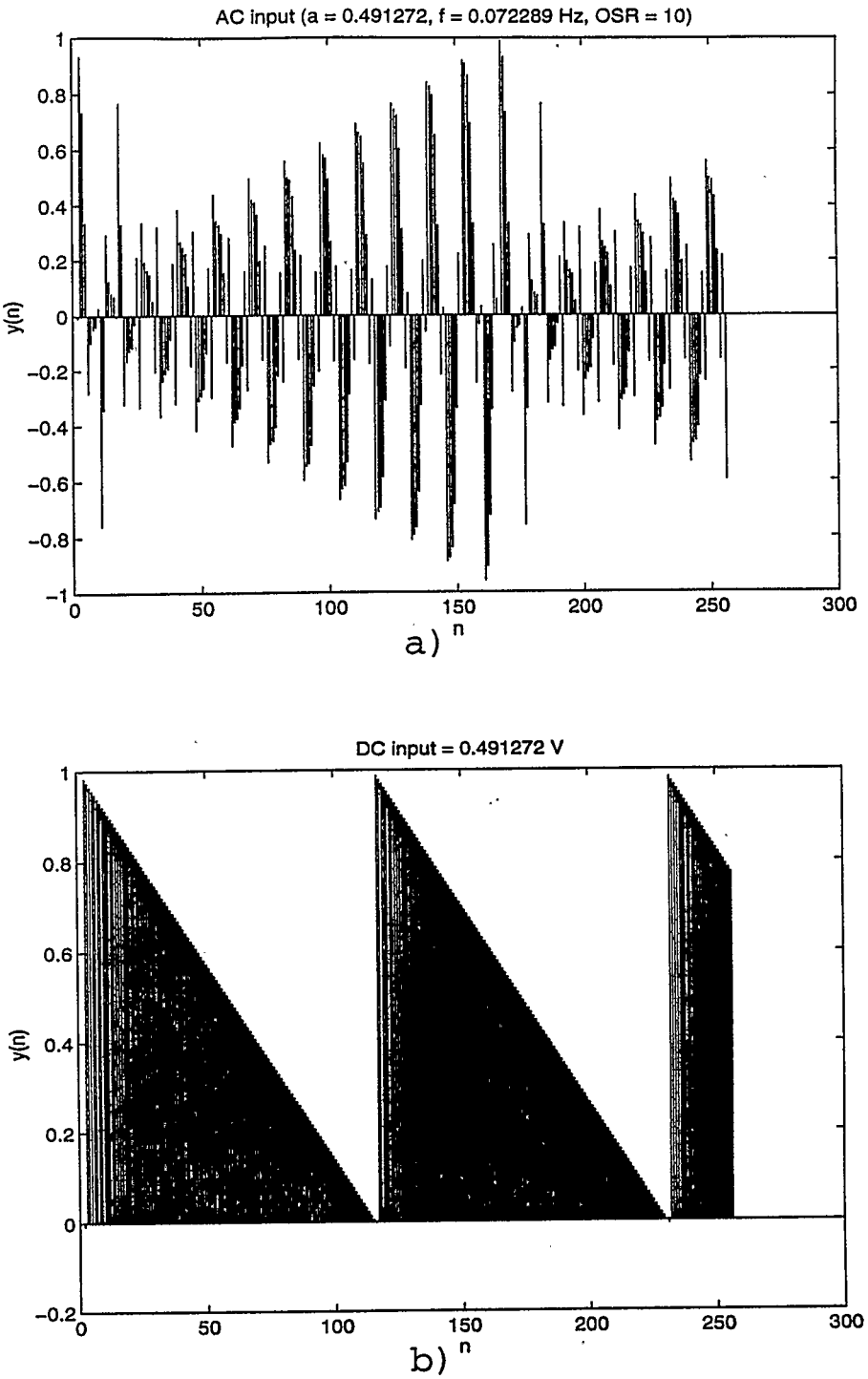
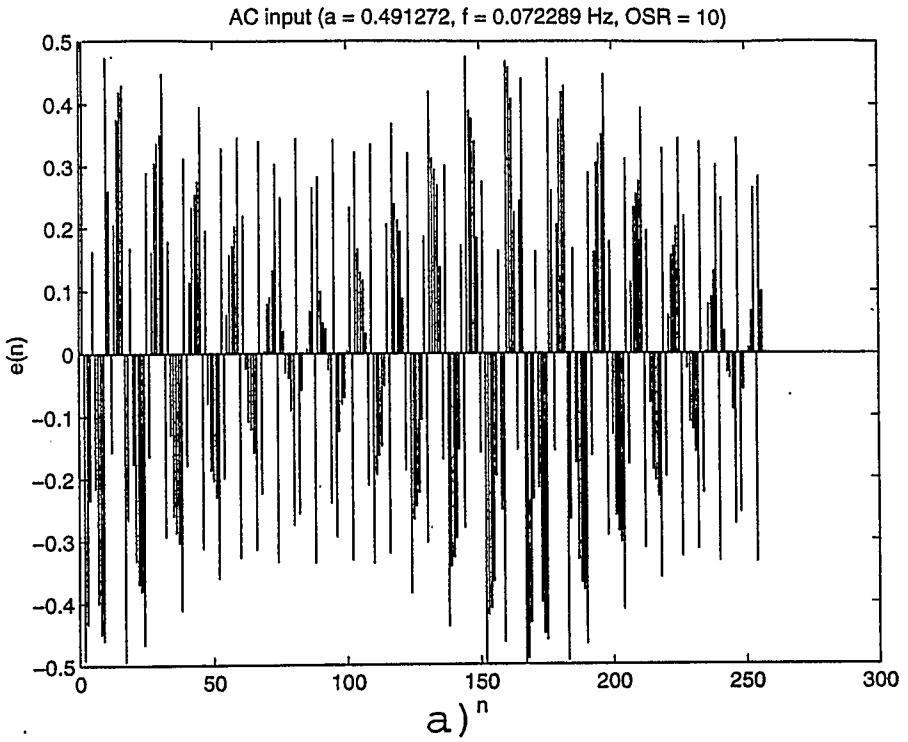


Figure 2.2. Quantizer input $y(n)$ for a single-loop converter with an a) AC input signal, b) DC input signal



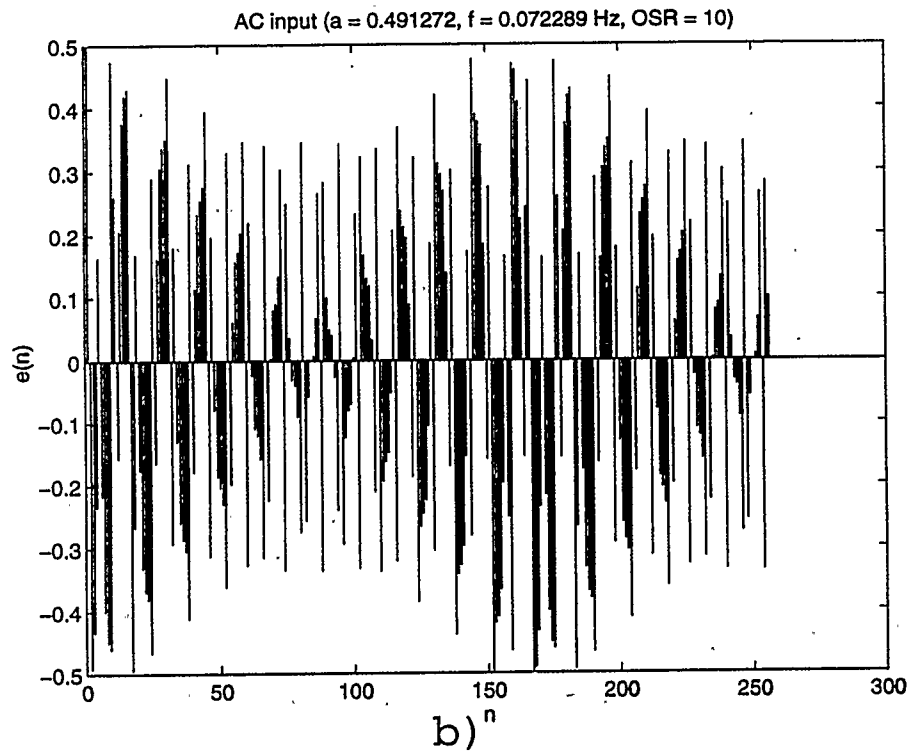


Figure 2.3. Quantization error for the single-loop converter (AC input) obtained by using a) difference equations b) closed-form solution

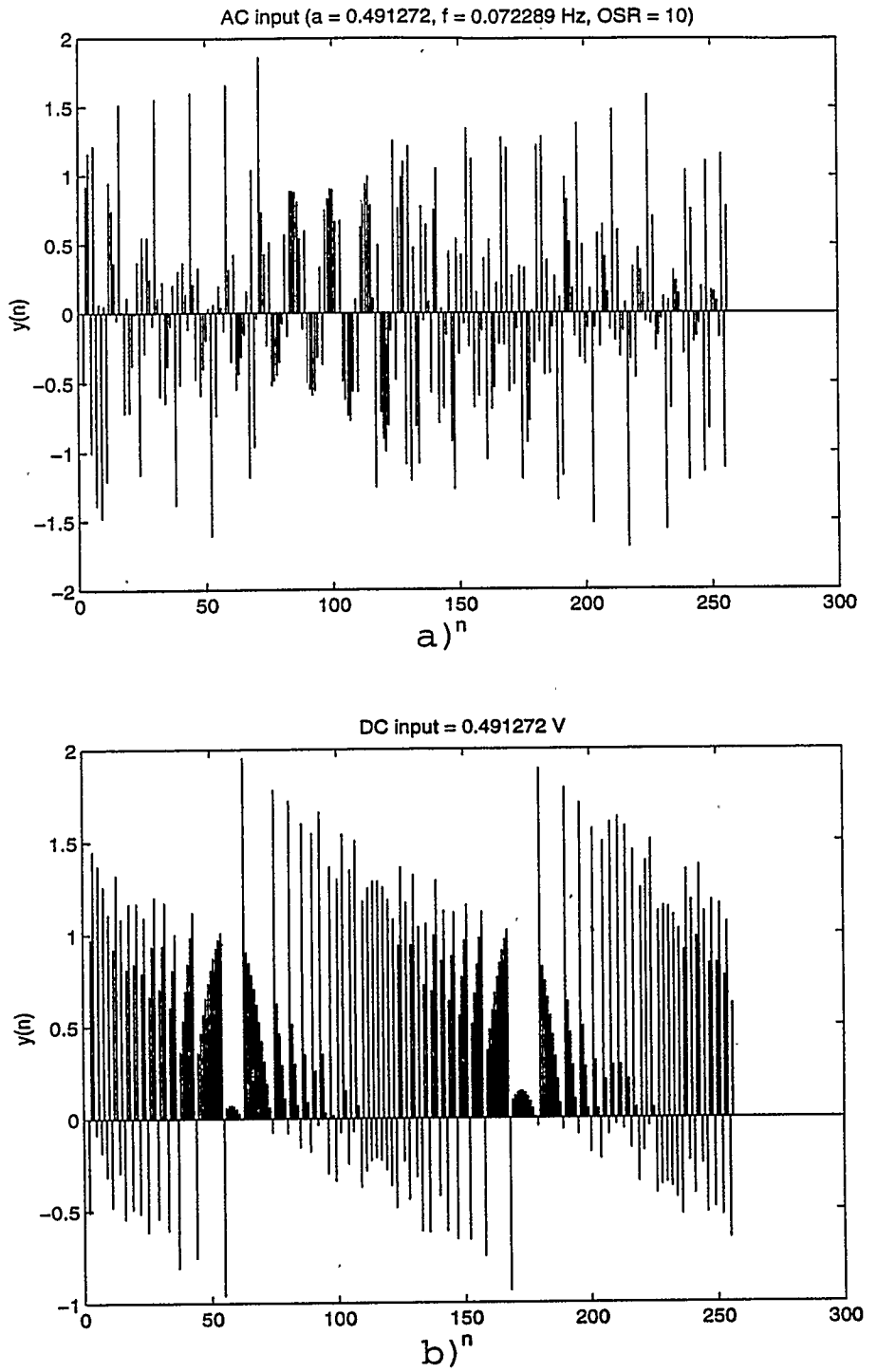
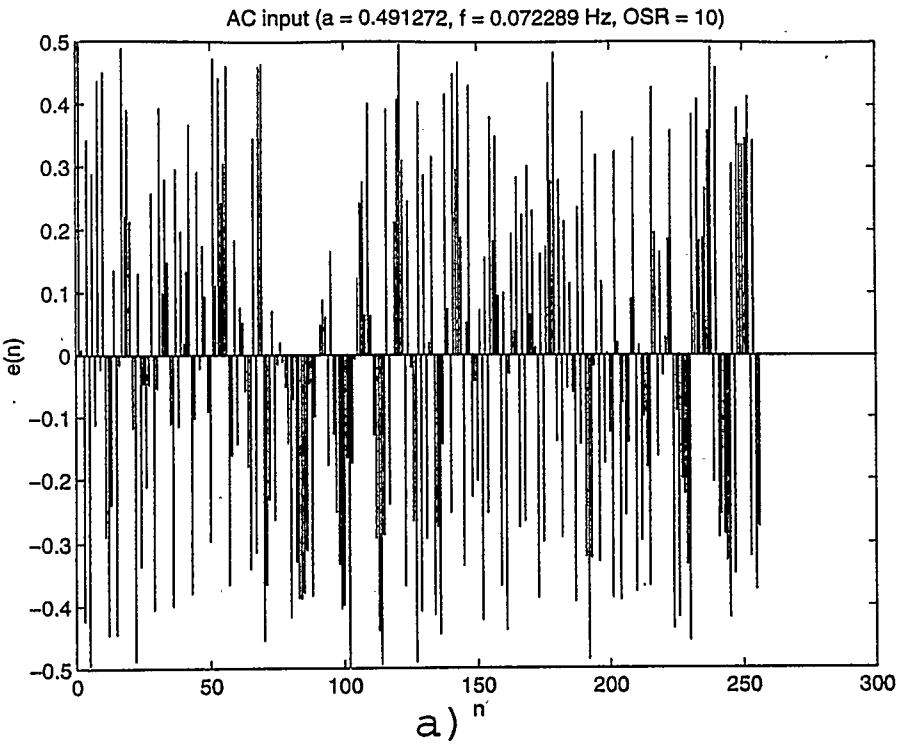


Figure 2.4. Quantizer input $y(n)$ for a double-loop converter with an a) AC input signal, b) DC input signal



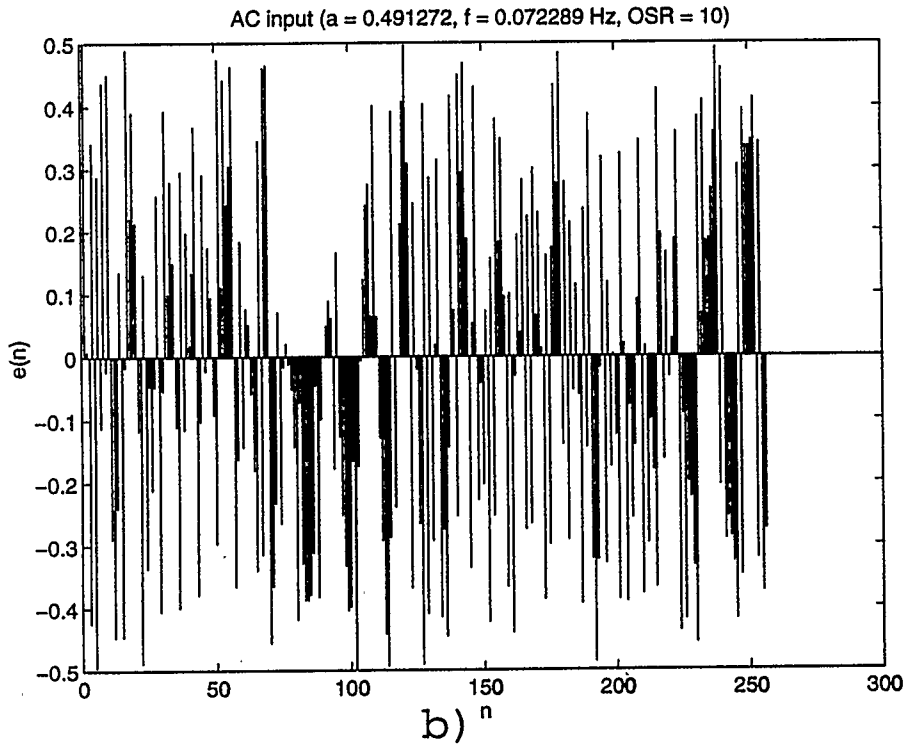


Figure 2.5. Quantization error for the double-loop converter (AC input) obtained by using a) difference equations b) closed-form solution

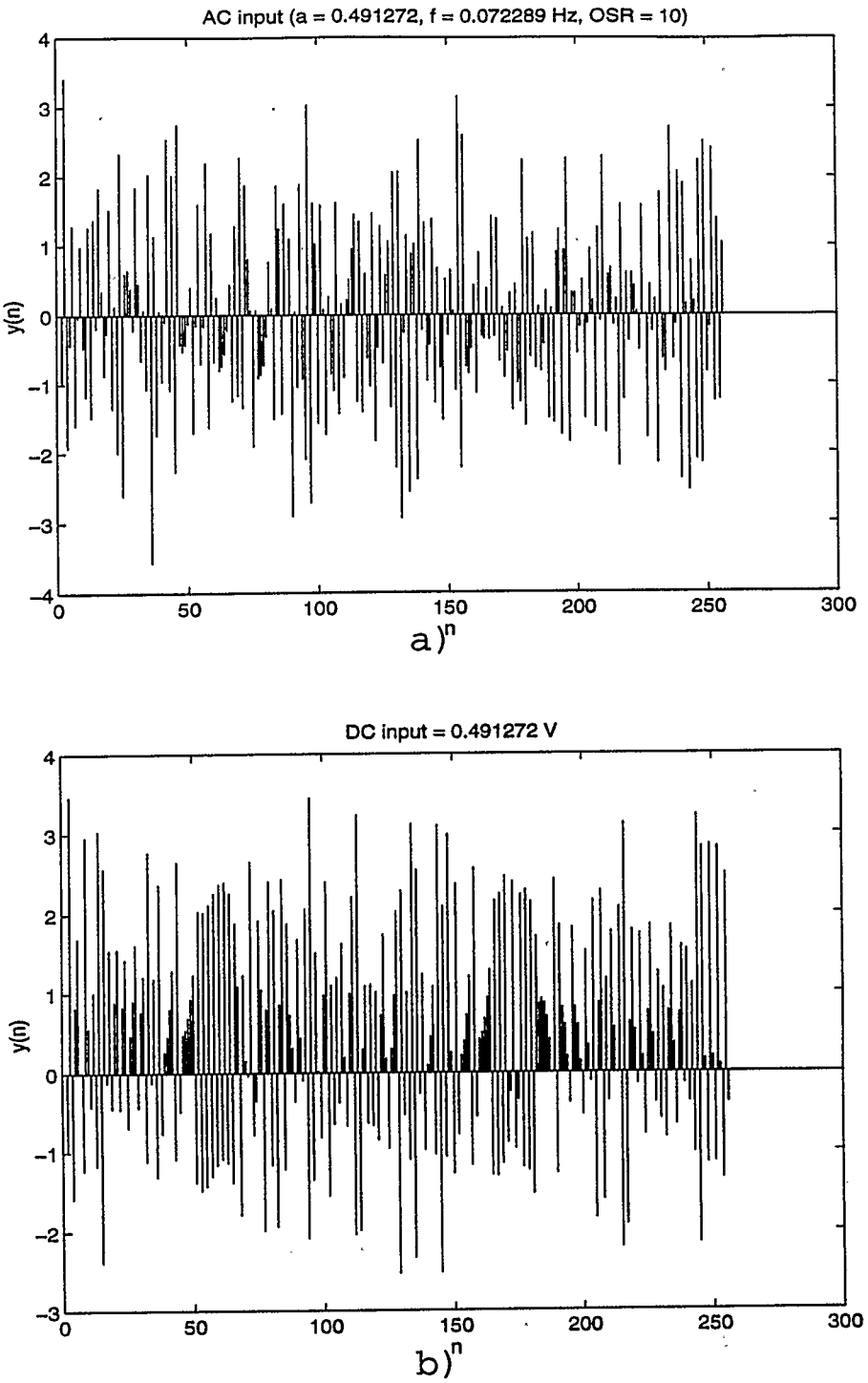
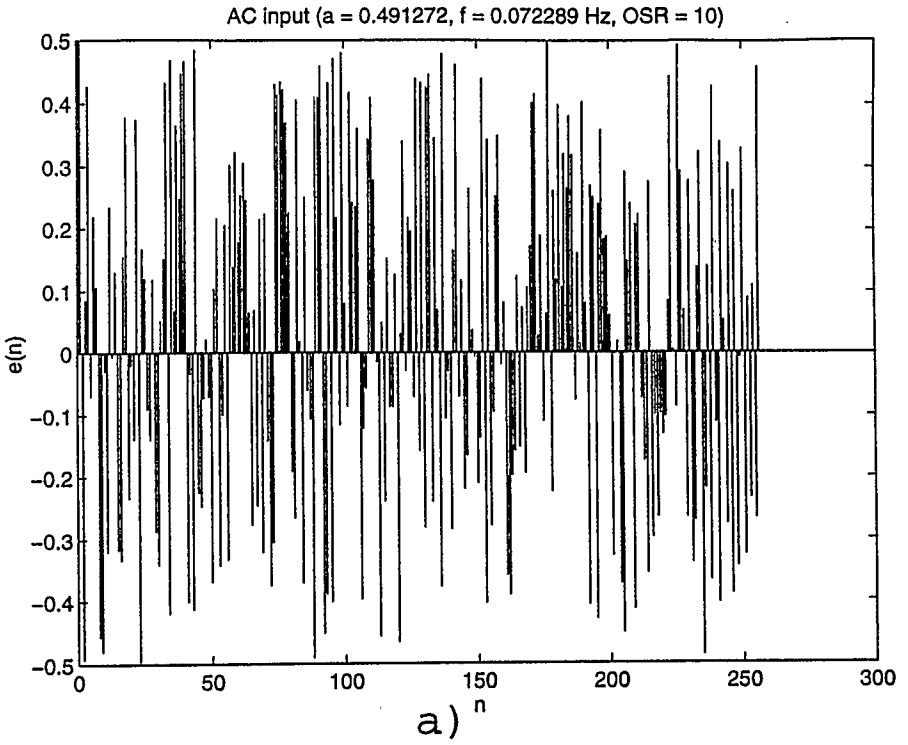


Figure 2.6. Quantizer input $y(n)$ for a triple-loop converter with an a) AC input, b) DC input



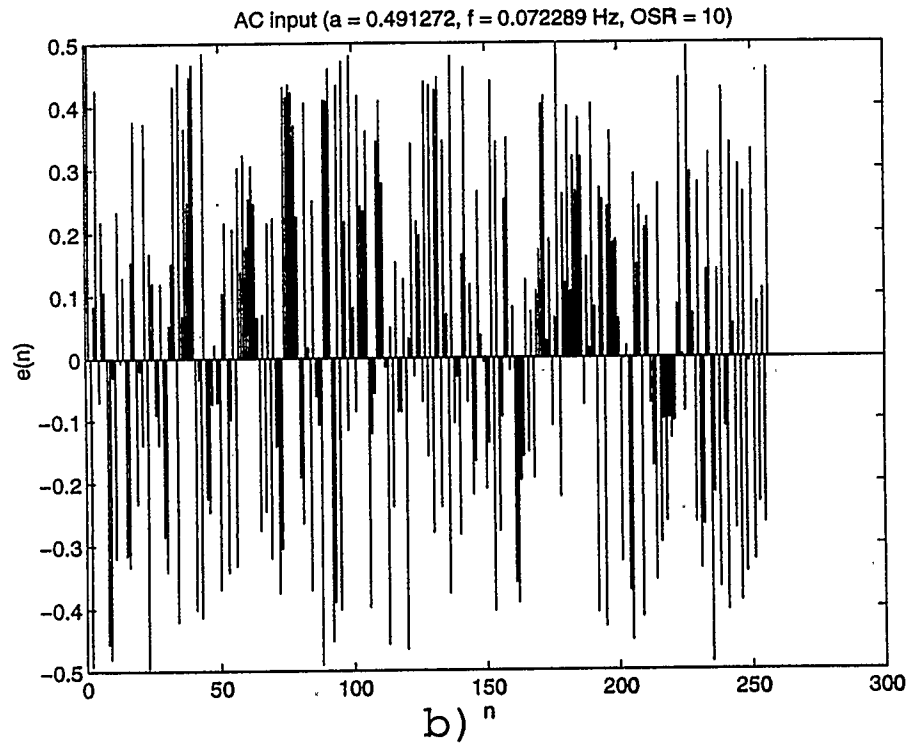


Figure 2.7. Quantization error for the triple-loop converter (AC input) obtained by using a)difference equations b) closed-form solution

CHAPTER 3

DERIVATION OF AN OPEN-LOOP EQUIVALENT SYSTEM FOR $\Sigma - \Delta$ CONVERTERS

3.1 Introduction

In the preceding chapter, the state-space formalism was exploited and applied to the derivation of the closed-form solution of the granular quantization error for a class of $\Sigma - \Delta$ converter configurations. In the present chapter, the state-space formalism is used to derive the corresponding closed-form solution of the quantizer output signal. The latter closed-form solution is established by means of a mathematical theorem complete with a formal proof. The solution is verified through its application to two practical examples and through comparison with the corresponding results obtained by direct computation using Matlab simulations. Then, the state-space formalism is used to derive a second closed-form solution for the quantizer output signal, which is then subsequently combined with the closed-form solution of the granular quantization error $e(n)$ in (2.12) to develop an equivalent open-loop system for the $\Sigma - \Delta$ modulator configuration in Fig. 2.1. This solution is also verified through its application to two practical examples and through comparison with the corresponding results obtained by direct computation using Matlab. The stablity of $\Sigma - \Delta$ converters is then examined with regard to subsystem parameters and also input signal amplitude, with the utitlity of these discussed issues being shown through the presentation of some practical application examples.

3.2 Derivation of the closed-form solution of the quantizer output signal

In this section, a closed-form solution is derived for the quantizer output signal $q(n)$ in the general $\Sigma - \Delta$ converter configuration in Fig. 2.1. Recall the state-space formulation for the linear time-invariant subsystem $\mathcal{N}$.

$$x(n+1) = Ax(n) + B_1u(n) + B_2q(n) \tag{3.1}$$

$$y(n) = Cx(n) + Du(n) \tag{3.2}$$

Then, by solving the recursion in (3.1), and by replacing the result in (3.2), one obtains

$$y(n) = CA^nx(0) + C\sum_{i=0}^{n-1} A^iB_1u(n-i-1) + C\sum_{i=0}^{n-1} A^iB_2q(n-i-1) + Du(n) \tag{3.3}$$

Let the quantizer operate in the no-overload mode, yielding

$$q(n) = \frac{\Delta}{2} + \Delta\lfloor\frac{1}{\Delta}y(n)\rfloor \tag{3.4}$$

Then, by substituting (3.3) into (3.4), one obtains

$$\begin{aligned} q(n) = \frac{\Delta}{2} + \Delta\left\lfloor\frac{1}{\Delta}\left[CA^nx(0) + C\sum_{i=0}^{n-1} A^iB_1u(n-i-1)\right.\right. \\ \left.\left. +C\sum_{i=0}^{n-1} A^iB_2q(n-i-1) + Du(n)\right]\right\rfloor \end{aligned} \tag{3.5}$$

Theorem 3.1 *If the scalars $C_{1\times N}(A+B_2C)^j_{N\times N}B_{2N\times 1}$ are integral numbers for each $j = 0, 1, \ldots, \infty$, then the closed form solution of (3.5) is given by*

$$\begin{aligned} q(n) = \frac{\Delta}{2} + \Delta\left[\left\lfloor\frac{1}{\Delta}\left(CA^nx(0) + C\sum_{k=0}^{n-1} A^kB_1u(n-k-1) + C\sum_{k=0}^{n-1} A^kB_2\frac{\Delta}{2} + Du(n)\right)\right\rfloor\right. \\ + C\sum_{j=0}^{n-1}(A+B_2C)^jB_2\left\lfloor\frac{1}{\Delta}\left(CA^{n-j-1}x(0) + C\sum_{l=0}^{n-2-j} A^lB_1u(n-2-j-l)\right.\right. \\ \left.\left.\left. +C\sum_{l=0}^{n-2-j} A^lB_2\frac{\Delta}{2} + Du(n-1-j)\right)\right\rfloor\right] \end{aligned} \tag{3.6}$$

The proof of Theorem 3.1 will be given after establishing Lemmas 6 and 7 below.

Lemma 6 *For any integer $n \geq 0$, one has*

$$A^n + \sum_{i=0}^{n-1} A^i B_2 C (A + B_2 C)^{n-i-1} = (A + B_2 C)^n \tag{3.7}$$

Proof: By induction. The lemma is clearly valid for $n = 0$. Therefore, it is sufficient to show that if it holds true for $n = m$, then it is also valid for $n = m + 1$.

For $n = m + 1$, the left-hand side of (3.7) may be written as

$$\begin{aligned} & A^{m+1} + \sum_{i=0}^{m} A^i B_2 C (A + B_2 C)^{m-i} \\ & \quad = \left[A^m + \sum_{i=0}^{m-1} A^i B_2 C (A + B_2 C)^{m-i-1} \right] (A + B_2 C) \end{aligned} \tag{3.8}$$

But, by setting $n = m$ in (3.7), one has

$$A^m + \sum_{i=0}^{m-1} A^i B_2 C (A + B_2 C)^{m-i-1} = (A + B_2 C)^m \tag{3.9}$$

Then, by substituting (3.9) in (3.8), and simplifying the result, one arrives at

$$(A + B_2 C)^m (A + B_2 C) = (A + B_2 C)^{m+1} \tag{3.10}$$

which shows that the lemma also holds true for $n = m + 1$. □

Lemma 7 *For any integer $n \geq 1$, one can show that*

$$
\begin{aligned}
&\sum_{i=0}^{n-1} A^i B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{n-1-i}x(o) + C \sum_{k=0}^{n-2-i} A^k B_1 u(n-2-i-k) \right.\right. \\
&\quad \left.\left. + C \sum_{k=0}^{n-2-i} A^k B_2 \frac{\Delta}{2} + Du(n-1-i) \right] \right\rfloor \\
&\quad + \sum_{i=0}^{n-1} A^i B_2 C \sum_{j-0}^{n-2-i} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \left(CA^{n-2-j-i}x(0) \right.\right. \\
&\quad + C \sum_{l=0}^{n-3-j-i} A^l B_1 u(n-3-j-l-i) \\
&\quad \left.\left. + C \sum_{l=0}^{n-3-j-i} A^l B_2 \frac{\Delta}{2} + Du(n-2-j-i) \right) \right\rfloor \\
&= \sum_{j=0}^{n-1} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{n-1-j}x(0) + C \sum_{l=0}^{n-2-j} A^l B_1 u(n-2-j-l) \right.\right. \\
&\quad \left.\left. + C \sum_{l=0}^{n-2-j} A^l B_2 \frac{\Delta}{2} + Du(n-1-j) \right] \right\rfloor
\end{aligned}
\tag{3.11}
$$

Proof: By induction. The lemma clearly holds true for $n-1=0$. Therefore, it suffices to show that if it is valid for $n=m$, then it also holds true for $n=m+1$.

For $n = m + 1$, the left-hand side of (3.11) can be written in the form

$$\begin{aligned}
&\sum_{i=0}^{m+1} A^i B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m+1-i}x(0) + C\sum_{k=0}^{m-i} A^k B_1 u(m-i-k) + C\sum_{k=0}^{m-i} A^k B_2 \frac{\Delta}{2} + Du(m+1-i) \right] \right\rfloor \\
&\quad + \sum_{i=0}^{m+1} A^i B_2 C \sum_{j=0}^{m-i} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \Big(CA^{m-j-i}x(0) \right. \\
&\quad \left. + C\sum_{l=0}^{m-1-j-i} A^l B_1 u(m-1-j-l-i) + C\sum_{l=0}^{m-1-j-i} A^l B_2 \frac{\Delta}{2} + Du(m-j-i) \Big) \right\rfloor \\
&= \left(A^{m+1} + \sum_{i=0}^{m} A^i B_2 C (A+B_2C)^{m-i} \right) B_2 \left\lfloor \frac{1}{\Delta} [Cx(o) + Du(0)] \right\rfloor \\
&\quad + \left(\sum_{i=o}^{m} A^i B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-i}x(0) + C\sum_{k=0}^{m-1-i} A^k B_1 u(m-1-i-k) \right. \right. \right. \\
&\quad \left. \left. \left. + C\sum_{k=0}^{m-1-i} A^k B_2 \frac{\Delta}{2} + Du(m-i) \right] \right\rfloor \right. \\
&\quad \left. + C\sum_{i=0}^{m} A^i B_2 C \sum_{j=0}^{m-i-1} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-j-i-1}x(0) \right. \right. \right. \\
&\quad \left. \left. \left. + C\sum_{l=0}^{m-2-j-i} A^l B_1 u(m-2-j-l-i) + C\sum_{l=0}^{m-2-j-i} A^l B_2 \frac{\Delta}{2} + Du(m-j-i-1) \right] \right\rfloor \right) \qquad (3.12)
\end{aligned}$$

But, by setting $n = m$ in (3.11), one gets

$$\begin{aligned}
&\sum_{i=0}^{m} A^i B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-i}x(o) + C\sum_{k=0}^{m-1-i} A^k B_1 u(m-1-i-k) \right. \right. \\
&\quad \left. \left. + C\sum_{k=0}^{m-1-i} A^k B_2 \frac{\Delta}{2} + Du(m-i) \right] \right\rfloor \\
&\quad + \sum_{i=0}^{m} A^i B_2 C \sum_{j-0}^{m-1-i} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \Big(CA^{m-1-j-i}x(0) \right. \\
&\quad + C\sum_{l=0}^{m-2-j-i} A^l B_1 u(m-2-j-l-i) \\
&\quad \left. + C\sum_{l=0}^{m-2-j-i} A^l B_2 \frac{\Delta}{2} + Du(m-1-j-i) \Big) \right\rfloor \qquad (3.13) \\
&= \sum_{j=0}^{m} (A+B_2C)^j B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-j}x(0) + C\sum_{l=0}^{m-1-j} A^l B_1 u(m-1-j-l) \right. \right. \\
&\quad \left. \left. + C\sum_{l=0}^{m-1-j} A^l B_2 \frac{\Delta}{2} + Du(m-j) \right] \right\rfloor
\end{aligned}$$

Then, by making use of (3.13) and Lemma 6 for $n = m$, the right-hand side of (3.12) simplifies to

$$\begin{aligned}
&(A+B_2C)^{m+1}B_2\left\lfloor\frac{1}{\Delta}[Cx(0)+Du(0)]\right\rfloor \\
&\quad+\sum_{j=0}^{m}(A+B_2C)^jB_2\left\lfloor\frac{1}{\Delta}\left[CA^{m-j}x(0)+C\sum_{l=0}^{m-1-j}A^lB_1u(m-1-j-l)\right.\right. \\
&\quad\left.\left.+C\sum_{l=0}^{m-1-j}A^lB_2\frac{\Delta}{2}+Du(m-j)\right]\right\rfloor \\
&=\sum_{j=0}^{m+1}(A+B_2C)^jB_2\left\lfloor\frac{1}{\Delta}\left[CA^{m+1-j}x(0)+C\sum_{l=0}^{m-j}A^lB_1u(m-j-l)\right.\right. \\
&\quad\left.\left.+C\sum_{l=0}^{m-j}A^lB_2\frac{\Delta}{2}+Du(m+1-j)\right]\right\rfloor
\end{aligned} \tag{3.14}$$

which shows that the lemma is also valid for $n = m + 1$. □

Proof of Theorem 3.1. By induction. The theorem clearly is valid for $n = 0$. In this way, it is sufficient to show that if it is valid for $n = 0, 1, 2, \ldots, m$, then it also holds true for $n = m + 1$.

By setting $n = m + 1$ in (3.5), one has

$$\begin{aligned}
q(m+1)=\frac{\Delta}{2}+\Delta\left\lfloor\frac{1}{\Delta}\left[CA^{m+1}x(0)+C\sum_{i=0}^{m}A^iB_1u(m-i)\right.\right. \\
\left.\left.+C\sum_{i=0}^{m}A^iB_2q(m-i)+Du(m+1)\right]\right\rfloor
\end{aligned} \tag{3.15}$$

By invoking (3.6) for the terms $q(m-i)$ in (3.5), and by making use of the property

$$\lfloor r+K\lfloor s\rfloor\rfloor=\lfloor r\rfloor+K\lfloor s\rfloor \tag{3.16}$$

(for r and s being real numbers and K being an integer) in the result, one can obtain

$$\begin{aligned} q(n) = \frac{\Delta}{2} + \Delta \left\lfloor \frac{1}{\Delta} \left[CA^{m+1}x(0) + C\sum_{i=0}^{m} A^i B_1 u(m-i) + C\sum_{i=0}^{m} A^i B_2 \frac{\Delta}{2} + Du(m+1) \right] \right\rfloor \\ + \Delta C \sum_{i=0}^{m} A^i B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-i}x(0) + C \sum_{k=0}^{m-1-i} A^k B_1 u(m-1-i-k) \right. \right. \\ \left. \left. + C \sum_{k=0}^{m-1-i} A^k B_2 \frac{\Delta}{2} + Du(m-i) \right] \right\rfloor \\ + \Delta C \sum_{i=0}^{m} A^i B_2 C \sum_{j=0}^{m-1-i} (A + B_2 C)^j B_2 \left\lfloor \frac{1}{\Delta} \left[CA^{m-1-j-i}x(0) \right. \right. \\ \left. \left. + C \sum_{l=0}^{m-2-j-i} A^l B_1 u(m-2-j-l-i) + C \sum_{l=0}^{m-2-j-i} A^l B_2 \frac{\Delta}{2} + Du(m-1-j-i) \right] \right\rfloor \end{aligned} \tag{3.17}$$

By invoking Lemmas 6, and 7 in (3.17), and by simplifying the result, one arrives at

$$\begin{aligned} q(m+1) = \frac{\Delta}{2} + \Delta \left\lfloor \frac{1}{\Delta} \left[CA^{m+1}x(0) C \sum_{i=0}^{m} A^i B_1 u(m-i) + C\sum_{i=0}^{m} A^i B_2 \frac{\Delta}{2} + Du(m+1) \right] \right\rfloor \\ + \Delta C \sum_{j=0}^{m} (A + B_2 C)^j B_2 \left\lfloor \frac{1}{\Delta} \left(CA^{m-j}x(0) \right. \right. \\ \left. \left. + C \sum_{l=0}^{m-1-j} A^l B_1 u(m-1-j-l) + C \sum_{l=0}^{m-1-j} A^l B_2 \frac{\Delta}{2} + Du(m-j) \right) \right\rfloor \end{aligned} \tag{3.18}$$

which shows that the theorem is also valid for $n = m + 1$. □

3.3 Derivation of the equivalent open-loop system

In this section, a closed-form solution is derived for the quantizer output signal $q(n)$. The result is subsequently combined with the closed-form solution of the granular quantization error $e(n)$ in (2.12) to develop an equivalent open-loop system for the $\Sigma - \Delta$ modulator configuration in Fig. 2.1.

By solving (2.1) recursively for $x(n)$, and by substituting the result into (2.2), the

quantizer input signal $y(n)$ may be obtained as

$$\begin{aligned} y(n) = CA^n x(0) + C\sum_{i=0}^{n-1} A^i B_1 u(n-i-1) \\ + C\sum_{i=0}^{n-1} A^i B_2 q(n-i-1) + Du(n) \end{aligned} \tag{3.19}$$

But, from (2.4),

$$q(n) = y(n) + e(n) \tag{3.20}$$

Therefore, by substituting (3.19) into (3.20), one has

$$\begin{aligned} q(n) = CA^n x(0) + C\sum_{i=0}^{n-1} A^i B_1 u(n-i-1) \\ + C\sum_{i=0}^{n-1} A^i B_2 q(n-i-1) + Du(n) + e(n) \end{aligned} \tag{3.21}$$

Theorem 3.2 *The closed-form solution of the quantizer output signal $q(n)$ in the $\Sigma - \Delta$ modulator configuration in Fig. 2.1 can be obtained as*

$$\begin{aligned} q(n) = C(A + B_2C)^n x(0) \\ + C\sum_{i=0}^{n-1}(A + B_2c)^i(B_1 + B_2D)u(n-i-1) \\ + C\sum_{i=0}^{n-1}(A + B_2C)^i B_2 e(n-i-1) + Du(n) + e(n) \end{aligned} \tag{3.22}$$

via (2.12).

The proof of Theorem 3.2 will be given after stating the following two lemmas.

Lemma 8 *For any integer $n \geq 0$, it may be shown that*

$$A^n + \sum_{i=0}^{n-1} A^i B_2 C(A + B_2C)^{n-i-1} = (A + B_2C)^n \tag{3.23}$$

Proof: By induction. The lemma is clearly valid for $n = 0$. Therefore, it is sufficient to show that if it holds true for $n = m$, then it is also valid for $n = m + 1$.

For $n = m + 1$, the left-hand side of (3.23) may be written as

$$\begin{aligned} & A^{m+1} + \sum_{i=0}^{m} A^i B_2 C (A + B_2 C)^{m-i} \\ & \quad = \left[A^m + \sum_{i=0}^{m-1} A^i B_2 C (A + B_2 C)^{m-i-1} \right] (A + B_2 C) \end{aligned} \tag{3.24}$$

But, by setting $n = m$ in (3.23), one has

$$A^m + \sum_{i=0}^{m-1} A^i B_2 C (A + B_2 C)^{m-i-1} = (A + B_2 C)^m \tag{3.25}$$

Then, by substituting (3.25) in (3.24), and simplifying the result, one arrives at

$$(A + B_2 C)^m (A + B_2 C) = (A + B_2 C)^{m+1} \tag{3.26}$$

which shows that the lemma also holds true for $n = m + 1$. □

Lemma 9 *It may be shown that for any integer $n \geq 0$,*

$$\sum_{i=0}^{n} A^i + \sum_{i=0}^{n} A^i B_2 C \sum_{l=0}^{n-i-1} (A + B_2 C)^l = \sum_{i=0}^{n} (A + B_2 C)^i \tag{3.27}$$

Proof: By induction. The lemma is clearly valid for $n = 0$. Therefore, it is sufficient to show that if it holds true for $n = m$, then it is also valid for $n = m + 1$.

For $n = m + 1$, the left-hand side of (3.27) may be written as

$$\begin{aligned} & \sum_{i=0}^{m+1} A^i + \sum_{i=0}^{m+1} A^i B_2 C \sum_{l=0}^{m-i} (A + B_2 C)^l \\ & \quad = \left(A^{m+1} + \sum_{i=0}^{m} A^i B_2 C (A + B_2 C)^{m-i} \right) \\ & \qquad + \left[\sum_{i=0}^{m} A^i + \sum_{i=0}^{m} A^i B_2 C \sum_{l=0}^{m-i-1} (A + B_2 C)^l \right] \end{aligned} \tag{3.28}$$

But, by setting $n = m$ in (3.27), one obtains

$$\sum_{i=0}^{m} A^i + \sum_{i=0}^{m} A^i B_2 C \sum_{l=0}^{m-i-1} (A + B_2 C)^l = \sum_{i=0}^{m} (A + B_2 C)^i \tag{3.29}$$

Then, by making use of (3.29) and Lemma 9 for $n = m$, the right-hand side of (3.28) simplifies to

$$(A + B_2C)^{m+1} + \sum_{i=0}^{m}(A + B_2C)^i = \sum_{i=0}^{m+1}(A + B_2C)^i \tag{3.30}$$

which shows that the lemma is also valid for $n = m + 1$. □

Proof of Theorem 3.2. By induction. The theorem clearly holds true for $n = 0$. In this way, it is sufficient to show that if it is valid for $n = 0, 1, 2, \ldots, m$, then it also holds true for $n = m + 1$.

By setting $n = m + 1$ in (3.21), one has

$$\begin{aligned} q(m+1) = {} & CA^{m+1}x(0) + C\sum_{i=0}^{m} A^iB_1u(m-i) \\ & + C\sum_{i=0}^{m} A^iB_2q(m-i) + Du(m+1) + e(m+1) \end{aligned} \tag{3.31}$$

Then, by invoking (3.22) for $q(m-i)$ in (3.31), and by grouping terms, one obtains

$$\begin{aligned} q(m+1) = {} & \left\{CA^mx(0) + C\sum_{i=0}^{m} A^iB_2C(A + B_2C)^{m-i}x(0)\right\} \\ & + \left\{C\sum_{i=0}^{m} A^i(B_1 + B_2D)u(m-i) + C\sum_{i=0}^{m} A^iB_2C \right. \\ & \left. \times \sum_{l=0}^{m-i-1} (A + B_2C)^l(B_1 + B_2D)u(m-i-l-1)\right\} \\ & + \left\{C\sum_{i=0}^{m} A^iB_2e(m-i) \right. \\ & \left. + C\sum_{i=0}^{m} A^iB_2C\sum_{l=0}^{m-i}(A + B_2C)^lB_2e(m-i-l-1)\right\} \\ & + Du(m+1) + e(m+1) \end{aligned} \tag{3.32}$$

Finally, by invoking Lemmas 8 and 9 in (3.32), one arrives at

$$\begin{aligned} q(m+1) = {} & C(A + B_2C)^{m+1}x(0) \\ & + C\sum_{i=0}^{m}(A + B_2c)^i(B_1 + B_2D)u(m-i) \\ & + C\sum_{i=0}^{m}(A + B_2C)^iB_2e(m-i) + Du(m+1) + e(m+1) \end{aligned} \tag{3.33}$$

which shows that the theorem is also valid for $n = m + 1$. □

To proceed further, (3.22) is recast into the form

$$q(n) = C(A + B_2C)^n x(0) + \sum_{i=0}^{n} h_S(i)u(n-i) + \sum_{i=o}^{n} h_N(i)e(n-i) \tag{3.34}$$

where

$$h_S(n) = \begin{cases} D, & n = 0 \\ C(A + B_2C)^{n-1}(B_1 + B_2D), & n \geq 1 \end{cases} \tag{3.35}$$

is the impulse response associated with signal transmission, and where

$$h_N(n) = \begin{cases} 1, & n = 0 \\ C(A + B_2C)^{n-1}B_2, & n \geq 1 \end{cases} \tag{3.36}$$

represents the impulse response associated with the noise transmission through the $\Sigma - \Delta$ modulator configuration in Fig. 2.1.

By using (3.22) in combination with (2.12), one obtains the equivalent open-loop system shown in Fig. 3.1 for the $\Sigma - \Delta$ modulator configuration in Fig. 2.1.

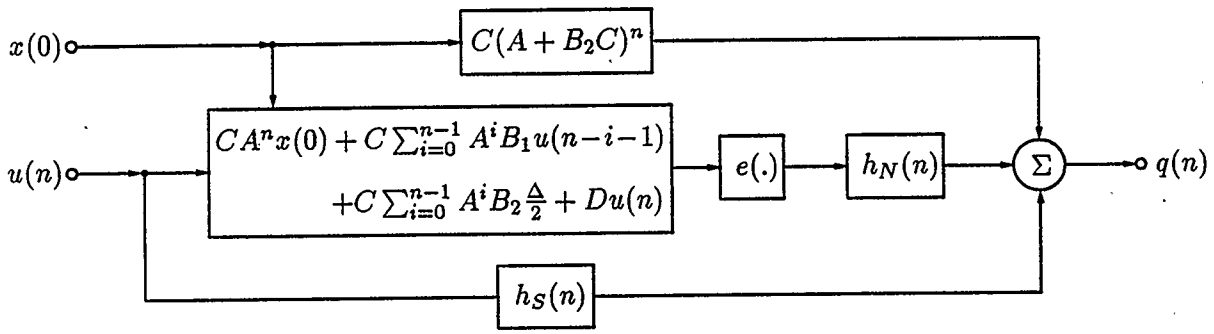


Figure 3.1. Equivalent open-loop system for the $\Sigma - \Delta$ modulator in Fig. 2.1

where the quantizer error sequence $e(n)$ is generated from the input sequence $u(n)$ according to the closed-form granular quantizer error equation. Therefore in the diagram above, the $e(\cdot)$ operator produces an output according to

$$e(x) = \frac{\Delta}{2} - \Delta\langle\frac{1}{\Delta}x\rangle \tag{3.37}$$

3.4 Application Examples

In this section, the above results are applied to the determination of the closed-form solution of the quantizer output $q(n)$ for conventional double-loop and triple-loop $\Sigma - \Delta$ converters. To simplify matters, it is assumed that $x(0) = 0$ throughout.

3.4.1 Double-loop $\Sigma - \Delta$ converter

For the double-loop converter in Fig. 1.9, by substituting for the matrix A, and the vectors B_1, B_2, C, and D from the state equations in (3.6), the closed-form solution for the quantizer output can be obtained as

$$q(n) = \frac{\Delta}{2} + \Delta\left[\left\lfloor\frac{1}{\Delta}\left(\sum_{i=0}^{n-1}\left[(i+1)u(n-i-1) - \frac{(i+2)\Delta}{2}\right]\right)\right\rfloor + \begin{bmatrix}0 & 1\end{bmatrix}\sum_{i=0}^{n-1}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}^i\begin{bmatrix}-1\\ -1\end{bmatrix}\left\lfloor\frac{1}{\Delta}\left(\sum_{l=0}^{n-2-i}\left[(l+1)u(n-2-i-l) - \frac{(l+2)\Delta}{2}\right]\right)\right\rfloor\right]$$

Similarly, by substituting for the matrix A, and the vectors B_1, B_2, C, and D in (2.12) and by using the open-loop equivalant system given in Fig. 3.1, the quantizer output may be obtained as

$$q(n) = z^{-1}u(n) + (1-z^{-1})^2\left(\frac{\Delta}{2} - \Delta\left\langle\frac{1}{\Delta}\left(\sum_{i=0}^{n-1}\left[(i+1)u(n-i-1) - \frac{(i+2)\Delta}{2}\right]\right)\right\rangle\right)$$

The quantizer output obtained using these two methods may be seen to be equivalent to that obtained using general difference equations (see Fig. 3.2).

3.4.2 Triple-loop $\Sigma - \Delta$ converter

For the triple-loop converter in Fig. 1.10, by substituting for the matrix A, and the vectors B_1, B_2, C, and D from the state equations in (3.6), the closed-form solution

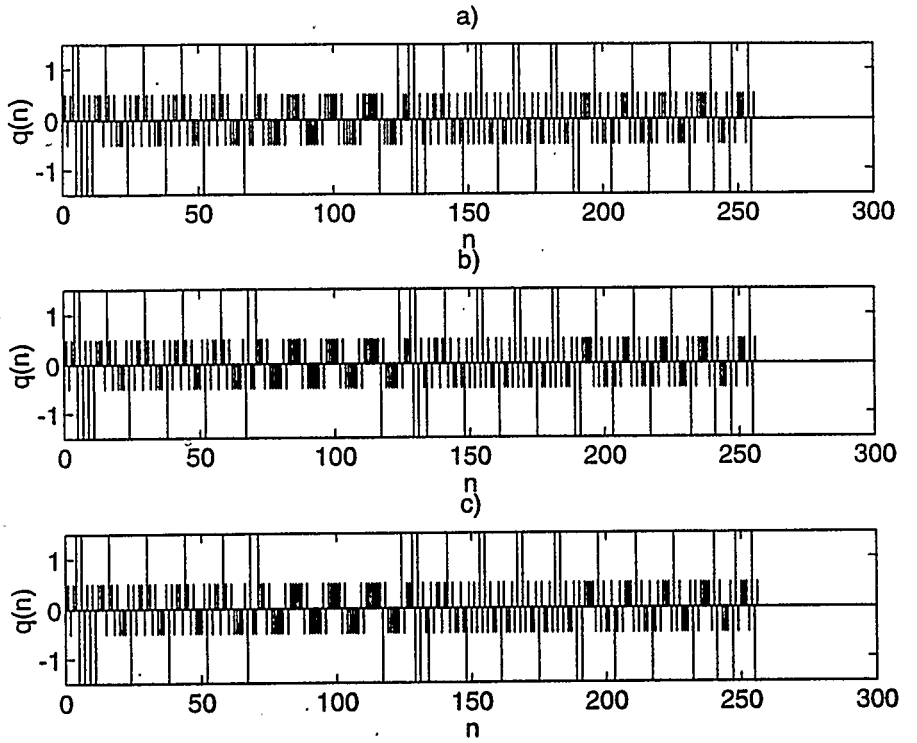


Figure 3.2. Quantizer output $q(n)$ for the double-loop converter obtained from a) difference equations, b) closed-form solution for $q(n)$, c) open-loop equivalent system output $q(n)$

for the quantizer output can be obtained as

$$q(n) = \frac{\Delta}{2} + \Delta \left[\left\lfloor \frac{1}{\Delta} \left(\sum_{i=0}^{n-1} \left[\frac{(i+1)(i+2)}{2} u(n-i-1) - \frac{(i+2)(i+3)\Delta}{4} \right] \right) \right\rfloor \right.$$
$$+ \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \sum_{i=0}^{n-1} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \\ 1 & 1 & -2 \end{bmatrix}^i \begin{bmatrix} -1 \\ -2 \\ -3 \end{bmatrix}$$
$$\left. \times \left\lfloor \frac{1}{\Delta} \left(\sum_{l=0}^{n-2-i} \left[\frac{(l+1)(l+2)}{2} u(n-2-i-l) - \frac{(l+2)(l+3)\Delta}{4} \right] \right) \right\rfloor \right]$$

Similarly, by substituting for the matrix A, and the vectors B_1, B_2, C, and D in (2.12) and by using the open-loop equivalant system given in Fig. 3.1, the quantizer output may be obtained as

$$q(n) = z^{-1}u(n)$$
$$+ (1-z^{-1})^3 \left(\frac{\Delta}{2} - \Delta \left\langle \frac{1}{\Delta} \left(\sum_{i=0}^{n-1} \left[\frac{(i+1)(i+2)}{2} u(n-i-1) - \frac{(i+2)(i+3)\Delta}{4} \right] \right) \right\rangle \right)$$

The quantizer output obtained using these two methods may be seen in Fig. 3.3.

3.5 Stability analysis of $\Sigma - \Delta$ converters

Stability is one of the primary concerns in the design of $\Sigma - \Delta$ A/D converters. The two causes of instability in these A/D converters may be traced to

1) Instability induced by the linear time-invariant subsystem.

2) Instability induced by the input signal amplitude.

The first is caused by improper design or by the required manufacturing precision for consituent converter configuration components not being met. In this case, the overall result is that the matrix $(A + B_2C)$ ends up with eigenvalues located outside the unit circle. Some discussion may be considered as to the preferred method used to analyse the stability of the single quantizer converter, as it is overall a nonlinear system which suggests the use of traditional nonlinear analysis techniques. In this fashion, this system may be examined with regard to

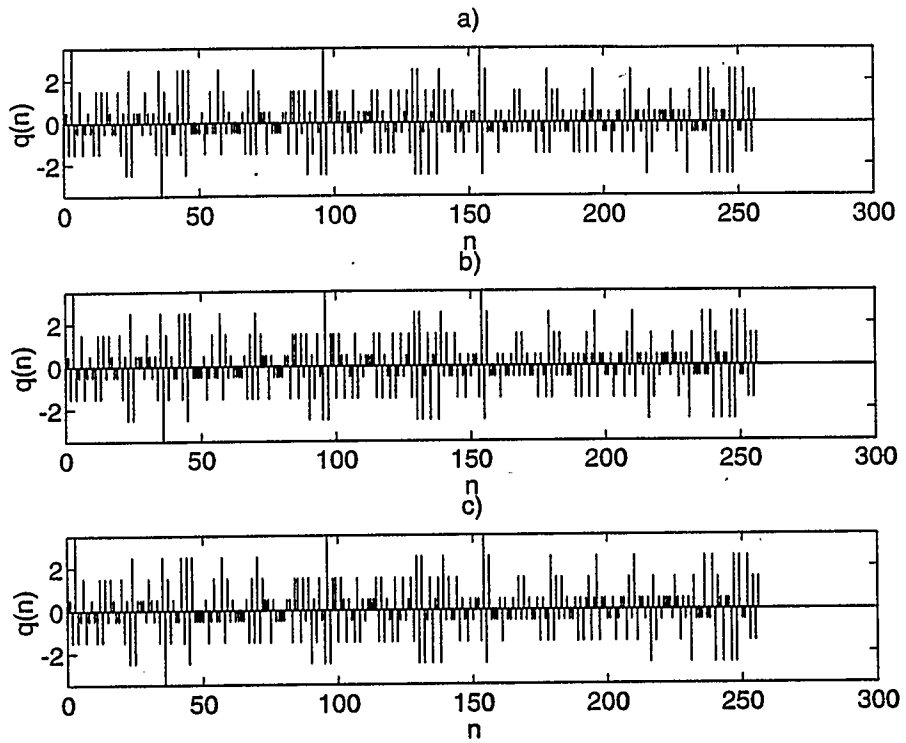


Figure 3.3. Quantizer output $q(n)$ for the triple-loop converter obtained from a) difference equations, b) closed-form solution for $q(n)$, c) open-loop equivalent system output $q(n)$

a) Zero input stablity.

b) Zero state stability.

as will be discussed in the following two subsections.

3.5.1 Zero input stablity

First let us examine the above equivalent system by using the concept of asymptotic stability.

The asymptotic stability concerns itself only with the state of the system, therefore imposing requirements on the motion of the state in state-space under *zero input* conditions only.

Definition 1 ([ZD63]) *A linear time-invariant discrete-time system described by (3.19) is said to be asymptotically stable (in the large) if*

i) For any $M > 0$, *there is a* $\delta > 0$ *such that* $\|x_0\| < \delta \Rightarrow \|x(n; x_0, 0; \mathbf{0})\| < \mathbf{M}$ *for* $n = 0, 1, 2, \ldots, \infty$.

ii) For all initial states x_0, $\lim_{n\to\infty} x(n; x_0, 0; \mathbf{0}) = \mathbf{0}$

where $x(n; x_0, 0; \mathbf{0})$ represents the state at time index n resulting from the initial state x_0 at time 0 and a zero input.

The above statements clearly imply that for zero input, the state remains bounded and tends to the zero state as $n \to \infty$. The implication of this result is that the multiplier values in the subsystem $\mathcal{N}$ must be chosen such that the matrix $(A + B_2C)$ has all eigenvalues located inside the unit circle for the system to be *zero input* stable.

3.5.2 Zero state stability

The second portion of this analysis will examine the *zero state* response of the equivalent system. Analysis of the quantized output (3.34) with regard to the zero state response only, reveals that it consists of the sum of two separate convolution sums as shown above. As the zero-state response consists of only the past and present

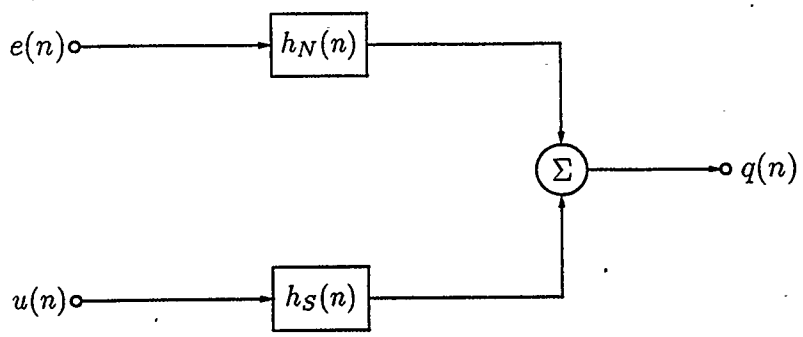


Figure 3.4. The zero state response of the linear time-invariant subsystem $\mathcal{N}$

values of the two inputs $u(n)$ and $e(n)$ given in (3.22), it is obviously causal (*non-anticipative*). Therefore, the two sequences $\{h_S(n)\}$ and $\{h_N(n)\}$ will be right-sided sequences, *i.e.* $h_S(n) = 0$ and $h_N(n) = 0$ for $n < 0$.

Now, from (3.35) it is clear that the two casual sequences $\{h_S(n)\}$ and $\{h_N(n)\}$ are related to the state matricies A, B_1, B_2, C, and D in the following manner

$$h_S(n) = \begin{cases} D, & n = 0 \\ C(A + B_2C)^{n-1}(B_1 + B_2D), & n \geq 1 \end{cases} \tag{3.38}$$

$$h_N(n) = \begin{cases} 1, & n = 0 \\ C(A + B_2C)^{n-1}B_2, & n \geq 1 \end{cases} \tag{3.39}$$

Given that the matrix $(A + B_2C)$ is stable, then the two responses will be absolutely summable. *i.e.*

$$\sum_{n=0}^{\infty} |h_S(n)| < \infty \tag{3.40}$$

and

$$\sum_{n=0}^{\infty} |h_N(n)| < \infty \tag{3.41}$$

As the impulse response and the transfer function are a z-transform pair, we may obtain the signal transfer function and the noise transfer function by taking the z-transforms of the two impulse responses.

$$\begin{aligned} \mathcal{Z}\{h_S(n)\} &= \sum_{n=0}^{\infty} h_S(n) z^{-n} \\ &= \sum_{n=1}^{\infty} C(A + B_2C)^{n-1}(B_1 + B_2D)z^{-n} + D \end{aligned} \tag{3.42}$$

and

$$\begin{aligned} \mathcal{Z}\{h_N(n)\} &= \sum_{n=0}^{\infty} h_N(n) z^{-n} \\ &= \sum_{n=1}^{\infty} C(A + B_2C)^{n-1}B_2 z^{-n} + 1 \end{aligned} \tag{3.43}$$

Simplification of the two above equations yields

$$\begin{aligned} H_S(z) &= C\left(\sum_{n=1}^{\infty}(A + B_2C)^{n-1}z^{-n}\right)(B_1 + B_2D) + D \\ &= C\left(z^{-1}\sum_{n=0}^{\infty}(z^{-1}(A + B_2C))^n\right)(B_1 + B_2D) + D \end{aligned} \tag{3.44}$$

and equivalently

$$H_N(z) = C\left(z^{-1}\sum_{n=0}^{\infty}(z^{-1}(A + B_2C))^n\right)B_2 + 1 \tag{3.45}$$

It is well known that for a matrix M whose eigenvalues lie inside the unit circle,

$$\sum_{n=0}^{\infty} M^n = (I - M)^{-1} \tag{3.46}$$

By using this matrix identity with $M = z^{-1}(A + B_2C)$, and by manipulating (3.44) and (3.45), one can obtain

$$H_S(z) = C(zI - (A + B_2C))^{-1}(B_1 + B_2D) + D \tag{3.47}$$

and

$$H_N(z) = C(zI - (A + B_2C))^{-1}B_2 + 1 \tag{3.48}$$

Due to the facts that the eigenvalues of M must lie within the unit circle (by assumption) and that z lies outside a circle whose radius is the magnitude of the largest eigenvalue of $(A + B_2C)$, the region of convergence of $H_S(z)$ and $H_N(z)$ will lie outside the outermost pole (due to causality) and will include the unit circle (due to stability). This result again implies that for bounded inputs $u(n)$ and $e(n)$, the zero state stability of the equivalent system requires that all multiplier values of the subsystem $\mathcal{N}$ be chosen such that the matrix $(A + B_2C)$ has all its eigenvalues located inside the unit circle.

3.6 Stability aspects of $\Sigma - \Delta$ converters due to input signal range

In the analysis of $\Sigma - \Delta$ converters, a complete analysis of the converter stability should always also reflect the dependence of the nonlinear subsystem on its input signal and not just stability due to system parameter values alone. Therefore, let us consider the conditions under which the quantizer is input signal or no-overload stable.

It was determined in Chapter 2 that for the quantizer to remain overload free, the input signal bound must be chosen such that $u(n) \in [-U_2, U_2]$. This would guarantee that the input signal to the quantizer would be bounded and the $\Sigma - \Delta$ converter remain free of limit cycles induced by quantizer saturation as documented in [Gra90]. In this manner, (2.35) or (2.43) may be used to define the no-overload stable region of the $\Sigma - \Delta$ converter as

$$U_2 \leq \frac{(M - \|C\| \cdot \|B_2\| W(1-\rho)^{-1}) \frac{\Delta}{2} - G}{(\|C\| \cdot \|(B_1 + B_2D)\| W(1-\rho)^{-1} + \|D\|)}$$

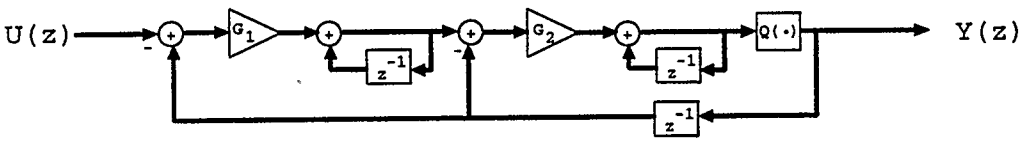
3.7 Application examples

In this section, the above results are applied to the stability analysis of conventional double-loop and triple-loop $\Sigma - \Delta$ converters. Throughout the discussions, it is assumed that $x(0) = 0$ to simplify matters.

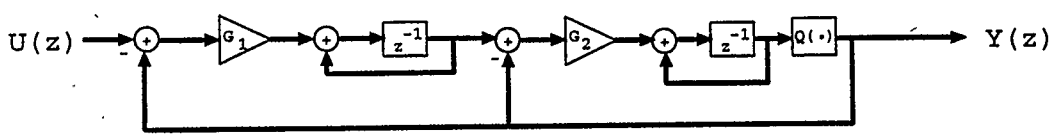
3.7.1 Double-loop $\Sigma - \Delta$ converters

In the design of higher order $\Sigma - \Delta$ converters realized as cascades of mutli-loop stages, the problem of stability arises. In particular, these multi-loop stages may become prone to limit cycle oscillation. For example, there are different variations of the double-loop converter in use which may be seen in the SOFOC1, and the SOFOC2 cascade converters seen in Figs. 1.15 and 1.16. The first version contains two simple integrators with a delay in the feedback path, and the second two delayed integrators as seen in Fig. 3.5.

The state-space equations for the two systems (for the case of matching gains



a) Double-loop converter (variation one)



b) Double-loop converter (variation two)

Figure 3.5. Two variations of the double-loop converter

$G_1 = G_2 = 1$) are

$$x(n+1) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \\ 1 & 1 & -2 \end{bmatrix} x(n) + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} w(n)$$

$$y(n) = \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} x(n) + \begin{bmatrix} 1 & 1 \end{bmatrix} w(n)$$

for the first double-loop converter, and

$$x(n+1) = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} x(n) + \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix} w(n)$$

$$y(n) = \begin{bmatrix} 0 & 1 \end{bmatrix} x(n) + \begin{bmatrix} 0 & 1 \end{bmatrix} w(n)$$

for the second double-loop converter, where the state vector $w(n)$ consists of the inputs $w_1 = U(z)$, and $w_2 = E_1(z)$. Using linear stability analysis shows that the eigenvalues of the matrix

$$(A + B_2C) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \\ 1 & 1 & -2 \end{bmatrix}$$

are

$$\lambda_1 = \lambda_2 = \lambda_3 = 0$$

and that the eigenvalues of the matrix

$$(A + B_2C) = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$$

are

$$\lambda_1 = \frac{1}{2} + j\frac{\sqrt{3}}{2}$$
$$\lambda_2 = \frac{1}{2} - j\frac{\sqrt{3}}{2}$$

Now it is well known that for a filter to be asymptotically stable, it must satisfy

$$\mu \max_i |\lambda_i| < 1$$

where μ is a positive real number greater than zero, and for the filter realization to have no limit cycles [BF77] it must satisfy

$$\underset{i}{max}|\lambda_i| < 1$$

It is quite obvious from the eigenvalues obtained, that the second converter does not meet the requirements of $\mu \underset{i}{max}|\lambda_i| < 1$ and is therefore prone to limit cycle oscillations. Stabilization of this converter is easily accomplished by the matrix transformation

$$A' = TAT^{-1}$$

This approach has been used to decrease the values of the two multipliers and thus move the eigenvalues of the matrix $(A + B_2C)$ into the stable region within the unit circle.

Similarly, it may be shown that a stable second order section of SOFOC1 may be driven into an unstable mode if the input signal amplitude exceeds the no-overload stable range of the converter given by $y(n) \leq 2$, determined in Section 2.6.2 (see Fig. 3.6).

3.7.2 Triple-loop $\Sigma - \Delta$ converter

There are different variations of the triple-loop converter [CT92], [BF94] as may be seen in Fig. 3.7.

The state-space equations for the systems (for the case of matching gains $G_1 = G_2 = G_3 = 1$) are

$$x(n+1) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix} x(n) + \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix} w(n)$$

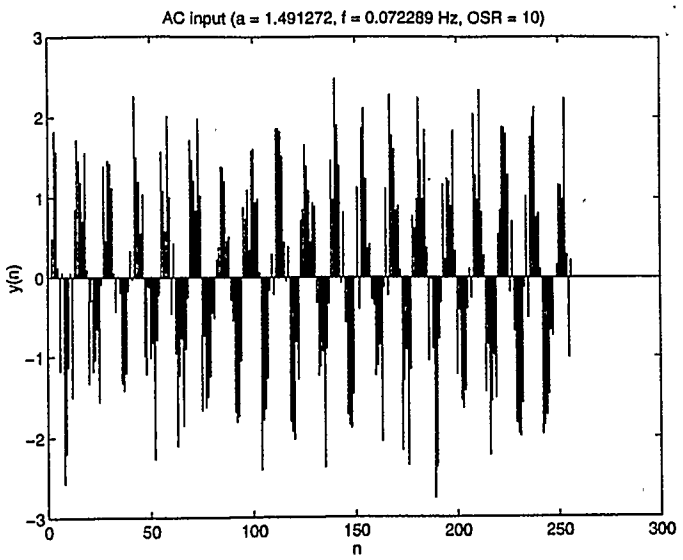


Figure 3.6. Input signal amplitude overloading the double-loop converter quantizer

$$y(n) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x(n) + \begin{bmatrix} 0 & 1 \end{bmatrix} w(n)$$

for the first triple-loop converter, and

$$x(n+1) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \\ 1 & 1 & -2 \end{bmatrix} x(n) + \begin{bmatrix} 1 & -1 \\ 1 & -2 \\ 1 & -3 \end{bmatrix} w(n)$$

$$y(n) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x(n) + \begin{bmatrix} 0 & 1 \end{bmatrix} w(n)$$

for the second triple-loop converter, where the state vector $w(n)$ consists of the inputs $w_1 = U(z)$, and $w_2 = E_1(z)$. Using linear stability analysis shows that the eigenvalues of the matrix

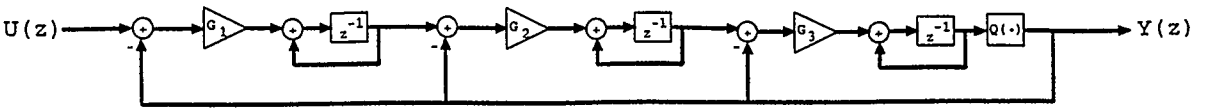
$$(A + B_2 C) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

are

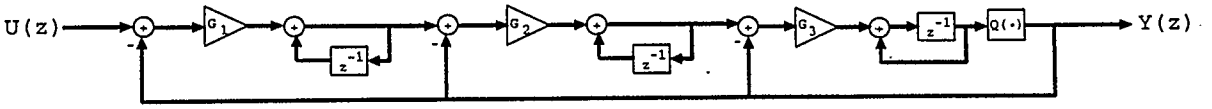
$$\lambda_1 = 0$$

$$\lambda_2 = 1 + j1$$

$$\lambda_3 = 1 - j1$$



a) Triple-loop converter (variation one)



b) Triple-loop converter (variation two)

Figure 3.7. Two variations of the triple-loop converter

and that the eigenvalues of the matrix

$$(A + B_2C) = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -2 \\ 1 & 1 & -2 \end{bmatrix}$$

are

$$\lambda_1 = \lambda_2 = \lambda_3 = 0$$

Stabilization of the first converter is again performed by the matrix transformation

$$A' = TAT^{-1}$$

to decrease the values of the three multipliers and thus move the eigenvalues of the matrix $(A + B_2C)$ into the stable region within the unit circle.

Similarly, it may be shown that a stable triple-loop converter may be driven into an unstable mode if the input signal amplitude exceeds the no-overload stable range of the converter given by $y(n) \leq 4$, determined in Section 2.6.3 (see Fig. 3.8).

3.8 Conclusion

This chapter has presented the derivation of a closed-form solution for the quantizer output as well as the derivation of an open-loop equivalent system based on the state-

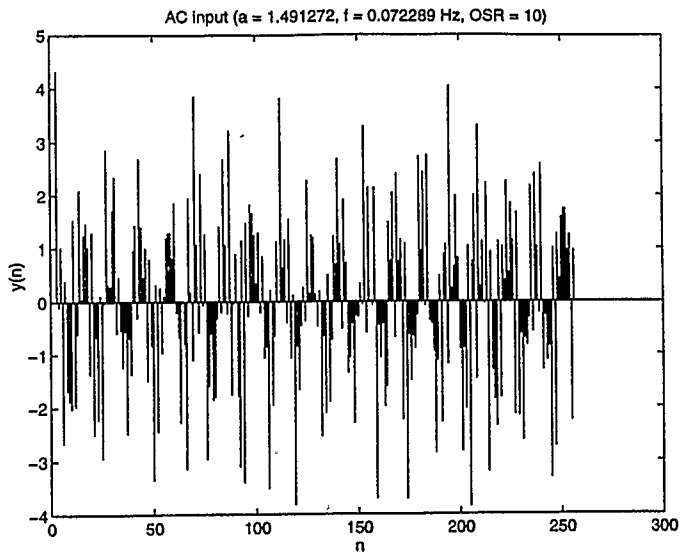


Figure 3.8. Input signal amplitude overloading the triple-loop converter quantizer

space equations describing a general class of single quantizer $\Sigma-\Delta$ converters. It then presented application examples verifying the closed-form solution of the quantizer output signal and the equivalent open-loop system, through the comparison with the corresponding results obtained by direct computation using Matlab. This chapter then examined stablity of $\Sigma - \Delta$ converters with regard to circuit parameters and also input signal amplitude. Finally, the utitlity of these discussed issues were then shown through the presentation of some application examples.

CHAPTER 4

SIGMA-DELTA CONVERTER QUANTIZATION NOISE SPECTRA

4.1 Introduction of spectral analysis methods used with $\Sigma-\Delta$ converters with a brief discussion of the linearized model

Spectral analysis is one of the most prominent tools used to characterize the behaviour and performance of $\Sigma - \Delta$ converters. There are several different techniques available for the determination of the spectral analysis of $\Sigma - \Delta$ converters. This chapter will examine the three most widely used spectral analysis methods for $\Sigma - \Delta$ converters, beginning with the linear system theory method (amounting to the representation of the quantization error by a white noise source), followed by the Fourier series analysis method, and ending with the characteristic function method. These three methods are later demonstrated through a practical application example. This chapter will then conclude with the use of the resulting spectral information in the calculation of the signal-to-quantization noise ratio which is the primary "yardstick" used to evaluate $\Sigma - \Delta$ converter resolution capability using all three methods.

4.2 Determination of spectral information using linear system theory

The primary technique used today for the analysis of $\Sigma - \Delta$ converters is based on a linear model for the converter [Can85]. In this technique, the quantizer error is replaced by an independent identically distributed noise source, allowing spectral

information to be derived using standard linear system theory. This method is based on the work of Bennett [Ben48] who derived his results under the conditions that

a) The quantizer has a large number of levels.

b) The quantizer bin-width is small.

c) The quantizer noise is signal independent (*i.e.* the quantizer error is uncorrelated to the quantizer input signal).

This approach is quite restrictive, however. This is due to the fact that single-loop $\Sigma - \Delta$ converters violate the conditions required for the above assumptions to hold, namely that the quantizer usually has very few levels (in most cases only two), the bin-width is not small, and that for simple inputs the quantizer error response is quite dependent on the amplitude and frequency of the input. Even though this linear model representation is incorrect for single-loop $\Sigma - \Delta$ converters, it does provide a good prediction for the behaviour of high-order $\Sigma - \Delta$ converters under most input conditions [HKB92], [WG90], and therefore must be included in this discussion of $\Sigma - \Delta$ quantization noise spectra for completeness.

Assuming that a one bit quantizer is utilized, the quantization error (e) added to the converter at the location of the quantizer will be bounded by $\pm(\Delta/2)$. If one also assumes that the quantizer error has a uniform probability over the interval $\left[\frac{-\Delta}{2}, \frac{\Delta}{2}\right]$, we may find its mean square value as [CT92]

$$\begin{aligned} e_{rms}^2 &= \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} e^2 de \\ &= \frac{\Delta^2}{12} \end{aligned} \tag{4.1}$$

This value will then be quite useful in the calculation of the spectral density of the sampled error (quantization error), and then the results obtained may be used to evaluate the system signal-to-quantization noise ratio.

The signal-to-quantization noise ratio may be calculated very simply for the single-loop $\Sigma - \Delta$ converter. When the quantized signal is sampled at $f_s = 1/T$, all the quantization error power will be folded into the range of frequencies over $0 \leq f \leq f_s/2$ [CB81]. Assuming that the quantization error is white, the spectral density of the sampled noise is found to be

$$E(f) = e_{rms}(2/f_s)^{1/2} = e_{rms}\sqrt{2T} \tag{4.2}$$

Therefore the spectral density of the output quantizer noise will be given by

$$N(f) = |1 - e^{-j\omega T}|E(f) \tag{4.3}$$

or

$$N(f) = e_{rms}\sqrt{2T}\left(sin\left(\frac{\omega T}{2}\right)\right)$$

In the search for more accurate methods in the analysis of the spectral behaviour, several alternative techniques have emerged. These alternatives have included an exact *Fourier* series representation of the error sequence [RL94], the method of characteristic functions [Gra89], [GCW89], [HKB92], and a continous time approximation [CB81]. With the development of a closed-form solution for the quantization error, it would be natural to exploit the techniques that utilize the closed-form solution of the granular quantization error to determine spectral information. The next two sections will discuss the determination of the quantization noise spectrum first by using a Fourier series representation of the closed-form solution of the quantization error and second by using the characteristic function method.

4.3 Determining spectral information using a direct Fourier series representation of the quantizer error sequence

The quantization noise spectrum is an important characteristic in evaluating the resolution performance of A/D converters. An exact determination of this noise

spectrum is made difficult in the case of oversampled $\Sigma - \Delta$ converters due to the constituent nonlinear coarse quantizer. One method that generally is used, is to determine the asymptotic autocorrelation of the quantization noise and then find the spectrum of this noise using the autocorrelation function and ergodic theory. A much more practical approach is to derive a Fourier series representation of the granular quantization error equation, such as used in the analysis of single-loop and double-loop converters based on the techniques originally used by Iwersen [Iwe69], or Clavier, Panter , and Grieg [CPG47]. This thesis will adopt the approach of Clavier, *et al.* as it allows the noise spectrum of the $\Sigma - \Delta$ quantizer to be determined using

$$\begin{aligned} e = e(u(n)) &= \sum_{l \neq 0} \frac{1}{2\pi jl} e^{2\pi jl \frac{u(n)}{\Delta}} \\ &= \sum_{l=1}^{\infty} \frac{1}{\pi l} \sin(2\pi l \frac{u(n)}{\Delta}) \end{aligned} \tag{4.4}$$

This Fourier series will hold for most converter input signals $u(n)$ provided that $u(n)$ satisfies the no-overload condition. Otherwise, $e(u(n))$ would not be a periodic function of $u(n)$ and the Fourier series representation could not be used. Similarly, one could write a Fourier series for e^2 as

$$\begin{aligned} e(u(n))^2 &= \frac{1}{12} + \sum_{l \neq 0} \frac{1}{2(\pi l)^2} e^{2\pi jl \frac{u(n)}{\Delta}} \\ &= \frac{1}{12} + \sum_{l=1}^{\infty} \frac{1}{2(\pi l)^2} \cos(2\pi l \frac{u(n)}{\Delta}) \end{aligned} \tag{4.5}$$

In this case, it is desired to study the behaviour of the normalized error sequence $\epsilon(n) = e(n)/\Delta$, this form of analysis may take two different paths. The first will be that of Clavier, *et al.* We know that the normalized error may be represented as

$$\begin{aligned} \epsilon(n) = \frac{e(n)}{\Delta} &= \frac{1}{2} - \left\langle \frac{1}{\Delta}[y(n)] \right\rangle \\ &= \frac{1}{2} - \left\langle \frac{1}{\Delta} \left[CA^n x(0) + C \sum_{i=0}^{n-1} A^i B_1 u(n-i-1) \right. \right. \\ &\quad \left. \left. + C \sum_{i=0}^{n-1} A^i B_2 \frac{\Delta}{2} + Du(n) \right] \right\rangle \end{aligned} \tag{4.6}$$

From (4.6) and (4.4) it may be immediately seen that

$$\begin{aligned}\epsilon(n) &= \sum_{l\neq 0} \frac{1}{2\pi jl} e^{2\pi jl\frac{y(n)}{\Delta}} \\ &= \sum_{l=1}^{\infty} \frac{1}{\pi l} \sin(2\pi l \frac{y(n)}{\Delta})\end{aligned} \tag{4.7}$$

Now for some specific examples of sequences $y(n)$, (4.7) can be easily used to obtain a direct Fourier series representation. Moments and spectral behaviour may then be determined with this representation, as the power spectral density (PSD) of the quantization error may be determined by

$$S_\epsilon(f) = \sum_{n=-\infty}^{\infty} X_n^2 \delta(f - nf_0) \tag{4.8}$$

where X_n represents the amplitude of the nth Fourier series component determined by using (4.7). One drawback of this method is that it may only work for input signals that are simple, *e.g.* sinusoids, as one may not be able to evaluate certain terms such as n^2 or n^3 in the exponential portion of (4.7). Another problem that can occur is that the ordinary Fourier series may not converge as $\epsilon(n)$ need not necessarily be a periodic function of n.

4.4 Determining spectral information using the characteristic function method

This approach will focus on the moments of the quantizer error process, leading to a variation of the characteristic function method described by Rice [Ric54], and Davenport and Root [DR58] under the name "transform method". In this case however, discrete time replaces continuous time, Fourier series replace Laplace transforms, and quasi-stationary processes will replace stationary processes as in [GCW89].

In this form of analysis, the main interest will be the long term average behaviour of the normalized error function $\epsilon(n)$. This will be in terms of the first moment,

second moment, and the autocorrelation functions defined by

$$M\{\epsilon(n)\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \epsilon(n)$$

$$M\{\epsilon(n)^2\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \epsilon(n)^2$$

$$r_\epsilon(k) = M\{\epsilon(n)\epsilon(n+k)\} = \lim_{N\to\infty} \frac{1}{n} \sum_{n=1}^{N} \epsilon(n)\epsilon(n+k)$$

respectively. As in the past [Gra90], a unified developement of both deterministic and random inputs will be performed using the technique of *quasi-stationary processes* proposed in [Lju87]. Using these results, the discrete time process $\epsilon(n)$ will be defined as quasi-stationary if there exists a finite constant C such that

$$E(\epsilon(n)) \leq C \quad ; \quad \text{for all} \quad n$$

$$|R_\epsilon(n,k)| \leq C \quad ; \quad \text{for all} \quad n,k$$

where $R_\epsilon(n,k) = E(\epsilon(n)\epsilon(k))$ and the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_\epsilon(n, n+k)$$

exists for each k. To avoid the implicit assumption of a zero mean, one must also include the first moment condition that the limit

$$\overline{m_\epsilon} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(\epsilon(n))$$

exists. Now given some process $w(n)$,

$$\overline{E}\{w(n)\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(w(n))$$

if the limit exists. Thus for a given quasi-stationary process $\{\epsilon(n)\}$, the autocorrelation will be given by

$$R_\epsilon(k) = \overline{E}\{\epsilon(n)\epsilon(n+k)\} \tag{4.9}$$

the mean by

$$m_\epsilon = \overline{E}\{\epsilon(n)\} \tag{4.10}$$

and the average power by

$$R_\epsilon(0) = \overline{E}\left\{\epsilon(n)^2\right\} \tag{4.11}$$

The power spectrum of the process will be defined as the discrete time Fourier transform of the autocorrelation function

$$S_\epsilon(f) = \sum_{n=-\infty}^{\infty} R_\epsilon(n) e^{-2\pi j f n}$$

where the frequency f is normalized to lie in $[0, 1]$.

Now by proceeding to invoke (4.32) and (4.4) to find the basic moments given by (4.9) - (4.11), one obtains the following expressions after some manipulation.

$$\begin{aligned} \overline{E}\{\epsilon(n)\} &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{l\neq 0} \frac{1}{2\pi j l} e^{2\pi j l \frac{y(n)}{\Delta}} \\ &= \sum_{l\neq 0} \frac{1}{2\pi j l} \overline{E}\left\{e^{2\pi j l \frac{y(n)}{\Delta}}\right\} \end{aligned} \tag{4.12}$$

$$\overline{E}\left\{\epsilon(n)^2\right\} = \frac{1}{12} + \sum_{l\neq 0} \frac{1}{2(\pi l)^2} \overline{E}\left\{e^{2\pi j l \frac{y(n)}{\Delta}}\right\} \tag{4.13}$$

and

$$R_\epsilon(k) = \sum_{i\neq 0} \sum_{l\neq 0} \frac{j}{2\pi i} \frac{j}{2\pi l} \overline{E}\left\{e^{2\pi j (i \frac{y(n)}{\Delta} + l \frac{y(n+k)}{\Delta})}\right\} \tag{4.14}$$

for $k \neq 0$. Redefining these results in terms of the one-dimensional and two-dimensional characteristic functions

$$\overline{\Phi_y}(l) = \overline{E}\left\{e^{2\pi j l \frac{y(n)}{\Delta}}\right\} \tag{4.15}$$

$$\Phi_y^{(k)}(i, l) = \overline{E}\left\{e^{2\pi j (i \frac{y(n)}{\Delta} + l \frac{y(n+k)}{\Delta})}\right\} \tag{4.16}$$

will give

$$\overline{E}\{\epsilon(n)\} = \sum_{l\neq 0} \frac{1}{2\pi j l} \overline{\Phi_y}(l) \tag{4.17}$$

$$\overline{E}\left\{\epsilon(n)^2\right\} = \frac{1}{12} + \sum_{l\neq 0} \frac{1}{2(\pi l)^2}\overline{\Phi_y}(l) \tag{4.18}$$

and

$$R_\epsilon(k) = -\sum_{i\neq 0}\sum_{l\neq 0} \frac{1}{2\pi i}\frac{1}{2\pi l}\Phi_{y(k)}(i,l) \tag{4.19}$$

for $k \neq 0$.

If the characteristic functions given by (4.15) and (4.16) can be evaluated (*i.e.* the appropriate moments of the quasi-stationary process are bounded), then using (4.17) - (4.19), the moments and spectrum of the quantizer error process can be computed.

4.5 Application Examples

The utility of the linear analysis method, and the two nonlinear analysis methods will now be demonstrated with the presentation of some application examples. In the following examples, the specific closed-form error equation will be evaluated under the conditon that $x(0) = 0$. This is justified, as it has been shown that the reset of intial states will not have any effect on the long term asymptotic behaviour of a $\Sigma - \Delta$ converter [Gra89].

4.5.1 The case of DC input signals

First we will examine the results obtained for DC input signals. The use of a DC input signal can represent a reasonable idealization to a slowly varying signal due to oversampling. The input to the system will be defined as

$$u(n) = X, \quad \text{for all} \quad n = 0, 1, 2, \ldots$$

where X is a constant amplitude. To evaluate the Fourier series using a general synthesis equation, it will be necessary to represent the two geometric matrix series

in (4.6) with their equivalent closed-form representations. It is well known that a general geometric series may be calculated using

$$\sum_{i=0}^{n-1} A^i = (I - A^{n+1})(I - A)^{-1}$$

A problem arises, however, if the matrix A contains an eigenvalue of unity. In this case, the geometric series may be evaluated as follows. Let the matrix A be replaced by the matrix $\acute{A}$ given by

$$\acute{A} = aA$$

where a represents some scalar. Then the series can be evaluated as

$$\begin{aligned}\sum_{i=0}^{n-1} A^i &= \lim_{a\to 1} \sum_{i=0}^{n-1} \acute{A}^i \\ &= \lim_{a\to 1} (I - a^{n+1}A^{n+1})(I - aA)^{-1}\end{aligned}$$

Next, let us represent the closed-form representation by some matrix P, where

$$P = \lim_{a\to 1} (I - a^{n+1}A^{n+1})(I - aA)^{-1}$$

Then using this result, the normalized error sequence may be represented as

$$\epsilon(n) = \frac{1}{2} - \left\langle \frac{1}{\Delta}\left(CPB_1X + CPB_2\frac{\Delta}{2}\right)\right\rangle \tag{4.20}$$

Evaluating (4.20) for the single-loop converter results in

$$\epsilon(n) = \frac{1}{2} - \left\langle \frac{nX}{\Delta} - \frac{n}{2}\right\rangle$$

Then using (4.5), the Fourier series representation may be written as

$$\epsilon(n) = \sum_{l\neq 0} \frac{1}{2\pi jl} e^{2\pi jl(\frac{nX}{\Delta} - \frac{n}{2})} = \sum_{l=1}^{\infty} \frac{1}{\pi l} \sin\left(2\pi l\left(\frac{nX}{\Delta} - \frac{n}{2}\right)\right)$$

and

$$\begin{aligned}\epsilon(n)^2 &= \frac{1}{12} + \sum_{l\neq 0} \frac{1}{2(\pi l)^2} e^{2\pi jl(\frac{nX}{\Delta} - \frac{n}{2})} \\ &= \frac{1}{12} + \sum_{l=1}^{\infty} \frac{1}{2(\pi l)^2} \cos\left(2\pi l\left(\frac{nX}{\Delta} - \frac{n}{2}\right)\right)\end{aligned}$$

These results may then be used to find moments and spectral behaviour.

Similarly, using the method of characteristic functions, one may obtain

$$\overline{E}\left\{\frac{1}{2}-\left\langle\frac{nX}{\Delta}-\frac{n}{2}\right\rangle\right\}=\sum_{l\neq 0}\frac{1}{2\pi jl}\overline{\Phi_y}(l)$$

$$\overline{E}\left\{\left(\frac{1}{2}-\left\langle\frac{nX}{\Delta}-\frac{n}{2}\right\rangle\right)^2\right\}=\frac{1}{12}+\sum_{l\neq 0}\frac{1}{2(\pi l)^2}\overline{\Phi_y}(l)$$

and

$$R_e(k)=-\sum_{i\neq 0}\sum_{l\neq 0}\frac{1}{2\pi i}\frac{1}{2\pi l}\Phi_y^{(k)}(i,l)$$

where

$$\overline{\Phi_y}(l)=\overline{E}\left\{e^{2\pi jl(\frac{nX}{\Delta}-\frac{n}{2})}\right\}$$

$$\Phi_y^{(k)}(i,l)=\overline{E}\left\{e^{2\pi j(i(\frac{nX}{\Delta}-\frac{n}{2})+l(\frac{nX}{\Delta}-\frac{n}{2}))}\right\}$$

4.5.2 The case of AC input signals

In a similar manner, results for (4.5), (4.6), and (4.17) - (4.19) may be obtained in the case of a general $\Sigma-\Delta$ converter for a sinsoidal input. Suppose that the input signal has the form

$$u(n)=\alpha\cos(n2\pi f/f_s+\theta)=\alpha\cos(n\omega+\theta)$$

where α, f, and θ are the amplitude, frequency, and phase, respectively, and f_s is the sampling frequency. This sampling frequency is assumed to be much larger than f since the the oversampling ratio, OSR $= f_s/2f$, is usually large. Therefore $\omega = 2\pi f/f_s$ will be small and $u(n)$ slowly varying. The amplitude $\alpha \leq U_1$ (due to

the no-overload condition) and the phase component will be set to zero for simplicity. In this case, the first matrix series may be evaluated as

$$\alpha \sum_{i=0}^{n-1} A^i \cos(n-i-1)\omega = \alpha \left(e^{j(n-1)\omega} \sum_{i=0}^{n-1} A^i e^{-ji\omega} + e^{-j(n-1)\omega} \sum_{i=0}^{n-1} A^i e^{ji\omega} \right) \tag{4.21}$$

Let the solution to the two geometric matrix series in (4.21) be represented by

$$R = \sum_{i=0}^{n-1} A^i e^{-ji\omega} = \left(I - e^{-j(n+1)\omega} A^{n+1} \right) \left(I - e^{-j\omega} A \right)^{-1} \tag{4.22}$$

and

$$S = \sum_{i=0}^{n-1} A^i e^{ji\omega} = \left(I - e^{j(n+1)\omega} A^{n+1} \right) \left(I - e^{j\omega} A \right)^{-1} \tag{4.23}$$

Therefore, substituting (4.22) and (4.23) into (4.21) yields

$$\alpha \sum_{i=0}^{n-1} A^i \cos(n-i-1)\omega = \alpha \left(e^{j(n-1)\omega} R + e^{-j(n-1)\omega} S \right)$$

Using this result, the normalized error sequence may be written as

$$\epsilon(n) = \frac{1}{2} - \left\langle \frac{\alpha}{\Delta} \left(e^{j(n-1)\omega} CRB_1 + e^{-j(n-1)\omega} CSB_2 \right) + CPB_2 \frac{1}{2} \right\rangle \tag{4.24}$$

Similarly, as for the DC input case, let us determine the Fourier series for the single-loop converter. From (4.24) the normalized error sequence is obtained as

$$\epsilon(n) = \frac{1}{2} - \left\langle \frac{\alpha}{\Delta} \frac{\cos(\frac{n-1}{2})\omega \sin\frac{n(-\omega)}{2}}{\sin(\frac{-\omega}{2})} - \frac{n}{2} \right\rangle$$

Then using (4.5), the Fourier series may be evaluated as

$$\epsilon(n) = \sum_{l \neq 0} \frac{1}{2\pi jl} \exp\left(2\pi jl \left\{ \frac{\alpha}{\Delta} \left(\frac{\cos(\frac{n-1}{2})\omega \sin\frac{n(-\omega)}{2}}{\sin(\frac{-\omega}{2})} \right) - \frac{n}{2} \right\} \right)$$

This form after manipulation and application of the *Jacobi-Anger* formula

$$e^{jz\sin\psi} = \sum_{m=-\infty}^{\infty} J_m(z) e^{jm\psi}$$

gives the result

$$\epsilon(n) = \sum_{m=-\infty}^{\infty} \sum_{l \neq 0} \frac{1}{2\pi jl} J_m(2\pi l\gamma) \times \exp\left(j\left[2\pi l\left\{\frac{\alpha}{2\Delta} - \frac{n}{2}\right\} + m\left(n - \frac{1}{2}\right)\omega\right]\right)$$

where J_m is the *Bessel* function of order m and

$$\gamma = \frac{\alpha}{2\Delta \sin(\frac{-\omega}{2})}$$

The square of the error sequence may also be calculated as

$$\epsilon(n)^2 = \frac{1}{12} + \sum_{l=1}^{\infty} \frac{1}{2(\pi l)^2} \cos\left(2\pi l\left[\frac{\alpha}{2\Delta} - \frac{n}{2} + \gamma \sin\left(n\omega - \frac{\omega}{2}\right)\right]\right)$$

These two results may then be used in the calculation of moments and spectral behaviour using (4.8).

In the same manner, one may use the method of characteristic functions to obtain

$$\overline{E}\left\{\frac{1}{2} - \left\langle \frac{\alpha}{\Delta} \frac{\cos(\frac{n-1}{2})\omega \sin\frac{n(-\omega)}{2}}{\sin(\frac{-\omega}{2})} - \frac{n}{2} \right\rangle\right\} = \sum_{l \neq 0} \frac{1}{2\pi jl} \overline{\Phi_y}(l)$$

$$\overline{E}\left\{\left(\frac{1}{2} - \left\langle \frac{\alpha}{\Delta} \frac{\cos(\frac{n-1}{2})\omega \sin\frac{n(-\omega)}{2}}{\sin(\frac{-\omega}{2})} - \frac{n}{2} \right\rangle\right)^2\right\} = \frac{1}{12} + \sum_{l \neq 0} \frac{1}{2(\pi l)^2} \overline{\Phi_y}(l)$$

and

$$R_\epsilon(k) = -\sum_{i \neq 0} \sum_{l \neq 0} \frac{1}{2\pi i} \frac{1}{2\pi l} \Phi_y^{(k)}(i, l)$$

where

$$\overline{\Phi_y}(l) = e^{j\pi \frac{a}{\Delta}} \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} e^{-j\pi ln} e^{j2\pi\gamma \sin(n\omega - \frac{\omega}{2})}$$

4.6 Using spectral information to calculate the signal-to-quantization noise ratio

The underlying reason for the determination of spectral behaviour of $\Sigma - \Delta$ converters has always been for the calculation of the achievable signal-to-quantization

noise ratio (SQNR) of the converter under investigation. The achievable SQNR in turn gives designers the resolution or performance capability of the converter.

In this way, by using results presented in the preceding sections, the power spectrum of the quantization error can be determined. The determination of the signal-quantization-noise ratio when using the linear analysis method may be accomplished by using (4.2) and (4.3). Thus, for the case of the single-loop converter, the quantization noise power in the signal band can then be evaluated by using

$$n_o^2 = \int_0^{f_0} |N(f)|^2 df \tag{4.25}$$

which gives the result

$$n_o^2 = 4e_{rms}^2(2T)\left[f - \frac{1}{(2\pi T)sin(2\pi fT)}\right] \tag{4.26}$$

Further manipulation and expansion of $sin2\pi T$ in a *Taylor's series* (truncating after the second term) will give the result

$$n_o^2 = e_{rms}^2\left(\frac{\pi}{3}\right)^2(2f_0T)^3 \tag{4.27}$$

where $2f_0T$ represents 1/(oversampling ratio) for the system. Finally the signal-to-quantization noise ratio may be calculated using the knowledge that the maximum input signal the converter may accomodate without saturating is $\Delta/2$ (due to a one-bit quantizer). This will give a signal power value of $(\Delta/(2\sqrt{2}))^2$ (assuming a sinsoidal input). Using previous results, the maximum signal-to-quantization noise ratio as a function of the oversampling ratio is found to be

$$\frac{s^2}{n_o^2} = \frac{4.5}{\pi^2}(2f_0T)^{-3} \tag{4.28}$$

Similarly, the power spectral density of the noise for a double-loop modulator may then be determined as

$$N(f) = 4e_{rms}\sqrt{2T}\left(sin\left(\frac{\pi T}{2}\right)\right)^2$$

The inband power may then be calculated as

$$n_o^2 = e_{rms}^2 \left(\frac{\pi^5}{5}\right)(2f_0T)^5 \tag{4.29}$$

resulting in a signal-to-quantization noise ratio of

$$\frac{s^2}{n_o^2} = \frac{7.5}{\pi^4}(2f_0T)^{-5} \tag{4.30}$$

This implies that without loss of generality the relative rms signal-to-quantization noise ratio of an Nth order modulator as a function of oversampling ratio may be calculated as

$$\frac{s^2}{n_o^2} = \frac{3(2N+1)}{2\pi^{2N}}(2f_0T)^{-(2N+1)} \tag{4.31}$$

(when the signal amplitude is maximum), and as

$$\frac{s^2}{n_o^2} = \frac{A^2}{\Delta^2}\frac{3(2N+1)}{2\pi^{2N}}(2f_0T)^{-(2N+1)} \tag{4.32}$$

when the signal amplitude satisfies $A < \Delta/2$.

For the cases of the two nonlinear analysis methods, the determination of the signal-to-quantization noise ratio may be accomplished by using (4.7) to find the amplitudes of the Fourier series components of the quantization error and substitute the results into the PSD equation for the Fourier series given by

$$S_\epsilon(f) = \sum_{n=-\infty}^{\infty} X_n^2 \delta(f - nf_0) \tag{4.33}$$

Similarly, the autocorrelation of the quantization error may be found using the method of characteristic functions as

$$R_\epsilon(k) = -\sum_{i\neq 0}\sum_{l\neq 0}\frac{1}{2\pi i}\frac{1}{2\pi l}\Phi_y^{(k)}(i,l) \tag{4.34}$$

The power spectrum of the error for both methods will be given by the discrete time Fourier transform of either (4.33) or (4.34). Using this information, the average

quantizer noise power at the output of the converter after subsampling or decimation may then be calculated as

$$\sigma_N^2 = \int_o^1 |H_N(f)|^2 \, |H_D(f)|^2 \, S_\epsilon(f) df \tag{4.35}$$

where $|H_N(f)|$ represents the magnitude response of the noise transfer function, and $|H_D(f)|$ the magnitude response of the decimation filter. Similarly the average signal power at the output after decimation may be calculated as

$$\sigma_S^2 = \int_0^1 |H_S(f)|^2 \, |H_D(f)|^2 \, S_S(f) df \tag{4.36}$$

where $|H_S(f)|$ and $|H_D(f)|$ represent the magnitude responses of the signal transfer function and the decimation filter respectively.

The signal-to-quantization noise ratio of the $\Sigma - \Delta$ converter may then be given by

$$SQNR = \frac{\sigma_S^2}{\sigma_N^2} \tag{4.37}$$

It may be observed that for $\Sigma - \Delta$ converters containing more than one loop, σ_N^2 is not a function of the input signal (it is approximately constant), and therefore the signal-to-quantization noise ratio will increase for an increase in the power of the input signal until the allowable input range has been exceeded and the quantizer is overloaded.

4.7 Conclusion

This chapter has examined the spectral analysis of $\Sigma - \Delta$ converters using three different methods based on

a) The linear white noise model for the quantizer error.

b) The Fourier series representation of the error sequence.

c) First and second moments of the error sequence derived using characteristic functions.

Some applications examples were then given and the three methods were used in the calculation of the converter signal-to-quantization noise ratio which is the primary method of determining $\Sigma - \Delta$ converter resolution.

CHAPTER 5

DESIGN OF $\Sigma - \Delta$ CONVERTERS

5.1 Introduction

This chapter will present a design methodology for the construction of cascaded sigma-delta converters which may be used in either Analog-to-Digital (A/D) or Digital-to-Analog (D/A) converters. Section 5.2 will formulate the overall design methodology in terms of the signal and noise transfer functions, while Section 5.3 will present various sigma-delta converters designed using such methods and the inherent benefits of such converters. Section 5.4 will then analyze the new converters with regard to their signal and noise transfer functions. Then the operation of these new converters will be characterized with regard to their signal-to-quantization noise ratio. This will be examined in subsections 5.5.1 and 5.5.2, the first with the signal-to-quantization noise ratio as a function of the input signal amplitude and the second with signal-to-quantization noise ratio as a function of the oversampling ratio.

5.2 Signal and noise transfer function properties necessary for $\Sigma - \Delta$ converters

It is clear from the literature [CT92] that the purpose of sucessive stages in cascade sigma-delta converter design is to further quantize the quantizer error from the first and sucessive stages (depending on the order of the system) for cancellation at the output. A more stringent criteria would be to state that the overall purpose of successive

stages is to cancel all sources of quantization noise but the last, where the quantization noise (error) will be shaped by the noise transfer function $H_N(z) = (1 - z^{-1})^N$ [BN94]. For baseband signals, the signal transfer function $H_S(z)$ is desired to have a lowpass magnitude/frequency response, while the noise (error) transfer function $H_N(z)$ is desired to have a highpass magnitude/frequency response. For all such desired system responses, the two final output signal spectra will be *complimentary*, with the noise signal occupying the portion of the frequency spectrum that may be filtered out leaving a final output signal (after decimation) that is contaminated with very little quantization noise.

It has been shown in the literature and discussed in Chapter 1, that through the use of cascaded stages, higher order sigma-delta converters may be constructed that have superior noise suppression in the desired frequency band of operation for lower oversampling ratios. This is due to the fact that the overall converter will retain the sum of the noise shaping properties of all the constiuent sections. For higher order converters consisting of the cascade of a second order and a first order section, one way to achieve the necessary cancellation (noise from the prior stages) is by isolating the quantization noise from the first section for processing and cancellation in the second section. The available internal signals that may be accessed for noise cancellation in the constituent second order section are its overall output signal

$$z^{-1}U(z) + (1 - z^{-1})^2 E_1(z) \tag{5.1}$$

the signal output from the first integrator

$$U(z) - (1 - z^{-1})E_1(z) \tag{5.2}$$

and the signal from the output of the second integrator (prior to the error source)

$$z^{-1}U(z) - z^{-1}E_1(z) - z^{-1}(1 - z^{-1})E_1(z) \tag{5.3}$$

as may be clearly seen in Fig. 5.1. It is most convenient to cancel out the contribution

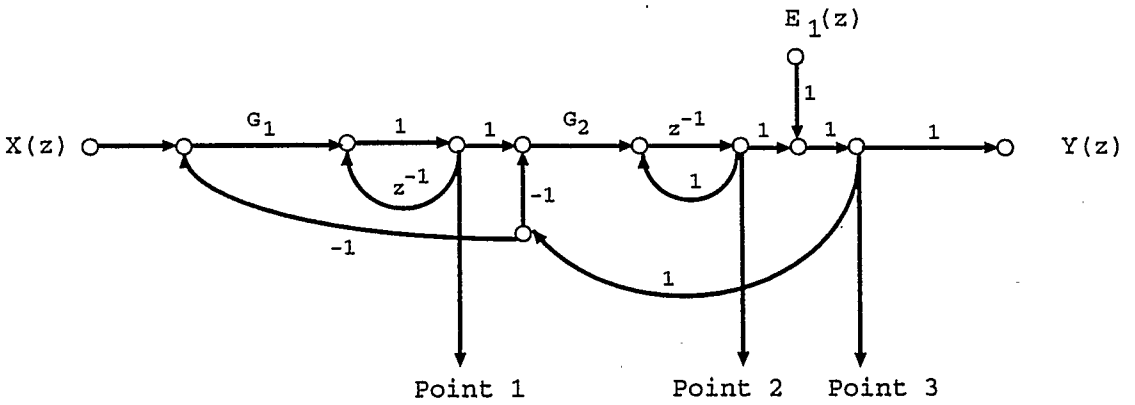


Figure 5.1. Available internal signals from the double-loop converter

of the input signal to the overall output formed through the second section. This may be achieved easily in two ways. The first is by subtracting a delayed version of (5.2) from (5.3), resulting in the negative delayed noise signal $-z^{-1}E_1(z)$. Alternatively, (5.1) may be subtracted from a delayed version of the signal in (5.2), resulting in the noise signal $-z^{-1}(1-z^{-1})E_1(z)$. This signal may then be passed through a discrete integrator to form the output $-z^{-1}E_1(z)$ which is the same result as obtained by using the first method. The third possible combination of signals (which is not considered), would be to subtract (5.3) from (5.1). This method is not chosen, as it would result in a noise signal $(1+2z^{-2})E_1(z)$ which could not be converted to the desired form $(-z^{-1}E_1(z))$ with simple structures as outlined for the other two methods outlined above. The isolated noise signal may then be fed into the second stage (a first-order sigma-delta converter) as the input signal. It will pass through this second stage being delayed and adding a second quantizer noise signal $(1-z^{-1})E_2(z)$. Cancellation is then achieved by making the noise transfer functions that shape the first noise source equivalent in both paths to the output. This will result in the two new sigma-delta

converter configurations as shown in Figs. 5.3 and 5.4 which shall be referred to as the Mash21a and Mash21b converters (as they are multistage noise shaping converters).

In a similar manner, a third-order sigma-delta converter may be constructed by cascading a first-order converter with a second-order converter. The available internal signals to be used for noise cancellation in the constituent first-order section are its output signal

$$z^{-1}U(z) + (1 - z^{-1})E_1(z) \tag{5.4}$$

and the output signal from the integrator

$$z^{-1}U(z) - z^{-1}E_1(z) \tag{5.5}$$

which are illustrated in Fig. 5.2.

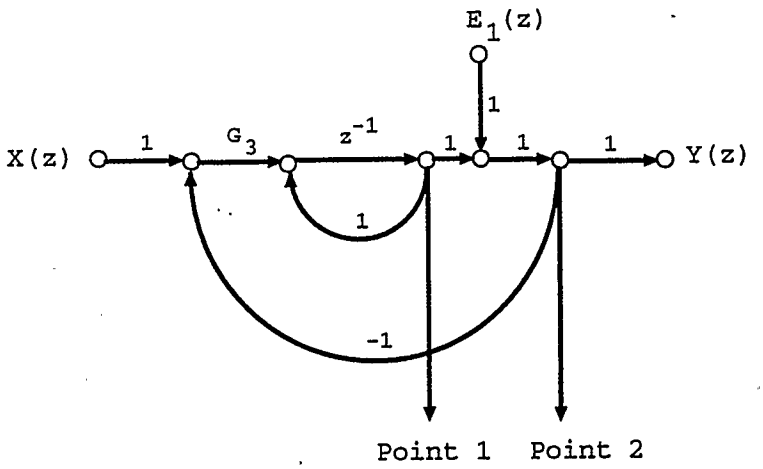


Figure 5.2. Available internal signals from the single-loop converter

As for the previous designs, it is expedient to cancel out the contribution of the input signal to the overall output formed from the second section. This is achieved by simply subtracting (5.6) from (5.7), resulting in the noise signal $-E_1(z)$. This

signal then serves as the input signal to the second stage and by making the transfer functions for the two paths to the output equal, this first noise source may be cancelled resulting in only the second noise signal $(1-z^{-1})^3 E_2(z)$ existing at the overall converter output. The converter implemented using this method may be seen in Fig. 5.5 and will be referred to as the Mash12 converter configuration.

5.3 Introduction of new cascade $\Sigma - \Delta$ converters

Several new cascade $\Sigma - \Delta$ converters were constructed using the techniques in the previous section, consisting of four third-order converters and one fourth-order converter. The third order converters will be presented first and consist of the Mash21a, Mash21b, Mash12, and Mash111 converters.

5.3.1 New third-order converters

The first two converters consist of the double-loop converter being the first stage in the design.The next one consists of the single-loop converter being the first stage which reduces the required amount of neccessary components by one delay and one adder. The final third-order converter utilizes a similar approach to that used by Candy and Temes [CT92] to obtain third-order noise shaping.

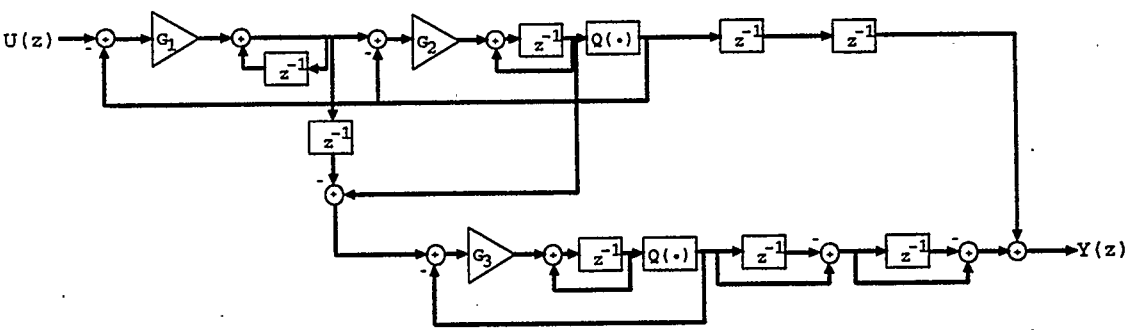


Figure 5.3. New Mash21a converter

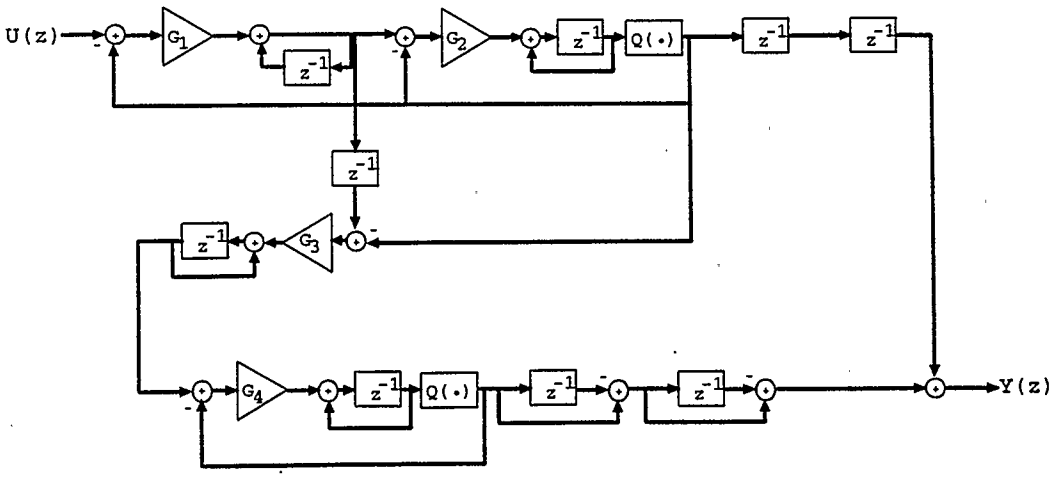


Figure 5.4. New Mash21b converter

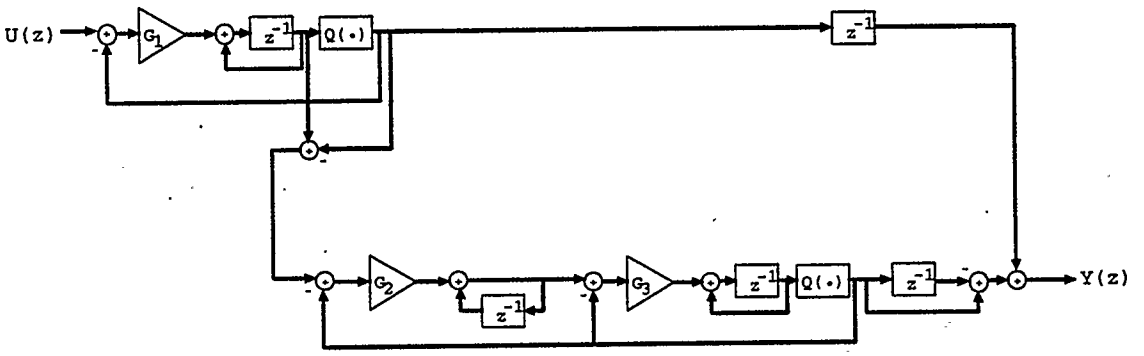


Figure 5.5. New Mash12 converter

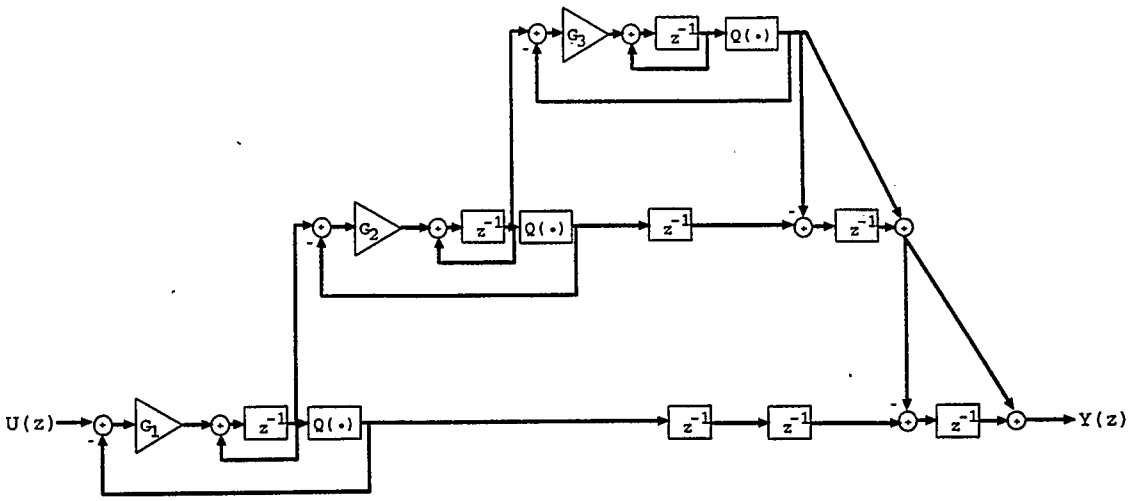


Figure 5.6. New Mash111 converter

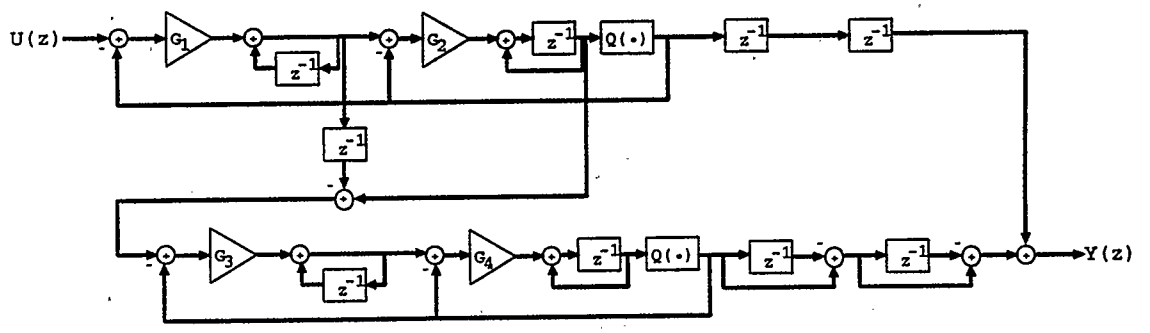


Figure 5.7. New Mash22 converter

5.3.2 A new fourth-order converter

The final new converter is a fourth-order converter using a cascade of two double-loop converters to provide fourth-order noise shaping.

5.4 Analysis of new converters with regard to signal and noise transfer functions

Verification of the obtained transfer functions of these new converters may be made via two methods, the first being signal flow-graph analysis and the second being state-space analysis. In the analysis of the signal and noise transfer functions to follow, the latter method will be employed. For the Mash21a converter one may write the state-space representation as

$$x(n+1) = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} x(n) + \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} w(n) \tag{5.6}$$

$$y(n) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 & -1 & -1 \end{bmatrix} x(n) + \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} w(n) \tag{5.7}$$

where $w(n) = \begin{bmatrix} u(n) & e_1(n) & e_2(n) \end{bmatrix}^t$. Substitution of the A, B, C, D matricies above in

$$H(z) = C[zI - A]^{-1}B + D \tag{5.8}$$

with simplification results in the following

$$Y(z) = z^{-3}U(z) + (1 - z^{-1})^3 E_2(z)$$

Similarly, the Mash21b converter may be represented as

$$x(n+1) = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} x(n) + \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} w(n) \tag{5.9}$$

$$y(n) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} w(n) \tag{5.10}$$

Substitution of the given A, B, C, D matricies in (5.8) with simplification results in

$$Y(z) = z^{-3}U(z) + (1 - z^{-1})^3 E_2(z) \tag{5.11}$$

In a like manner, the output equations for the Mash12 and Mash111 converters may be generated as

$$Y(z) = z^{-2}U(z) + (1 - z^{-1})^3 E_2(z) \tag{5.12}$$

$$Y(z) = z^{-3}U(z) + (1 - z^{-1})^3 E_2(z) \tag{5.13}$$

which clearly shows that the remaining quantizer error sequence is shaped by the desired third order noise transfer function $H_N(z) = (1 - z^{-1})^3$.

Writing the state-space eqautions for the Mash22 converter and then solving for them using (5.8), one obtains the output

$$Y(z) = z^{-3}U(z) + (1 - z^{-1})^4 E_2(z) \tag{5.14}$$

which has the signal transfer function $H_S(z) = z^{-3}$ and the desired fourth order noise shaping transfer function $H_N(z) = (1 - z^{-1})^4$.

Variants of each of these converter configurations were simulated, as it was found that the inclusion of additional delays in the system had the effect of further decorrelating the remaining error signal in the output from the input signal. This may be explained quite simply as the error sequence (generated by the first quantizer and correlated to the input signal) is further decoupled in relation to its nearest neighbor in the sequence by the inclusion of the additional delay. The second quantizer then operates on the additionally delayed error sequence from the first stage yielding a second error sequence that is in turn, further decorrelated with the original system input.

The benefit of the additional delay in both paths to the output may be clearly seen in the comparison between the noise output power spectral density of the Mash21a converter and a variant of the Mash21a (Fig. 5.8) seen in Figs. 5.9 and 5.10. Other

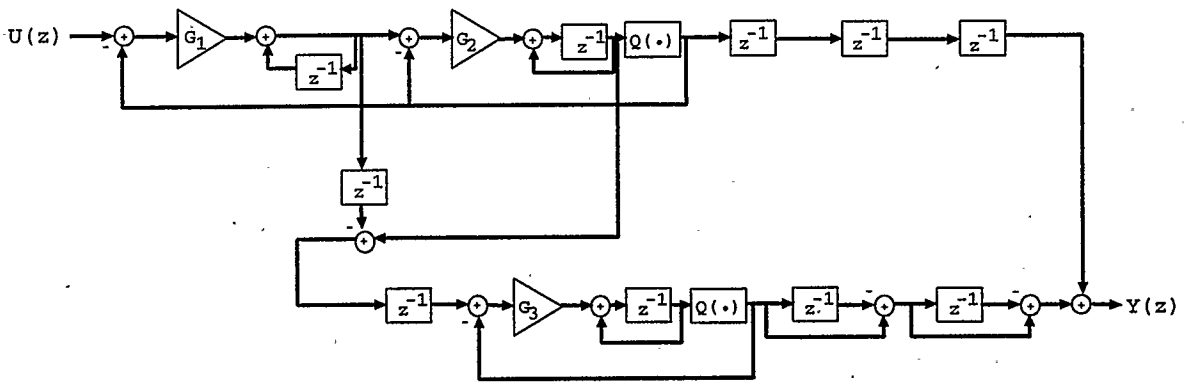


Figure 5.8. A variation of the Mash21a converter

such variants were tried and simulation results have shown further improvements in the signal-to-quantization noise ratio, but improvements were only significant with the addition of the first delay. The relation between further delay additions and improvements in the signal-to-quantization noise ratio decrease exponentially and would not be viable for the additional amount of hardware required for implementation.

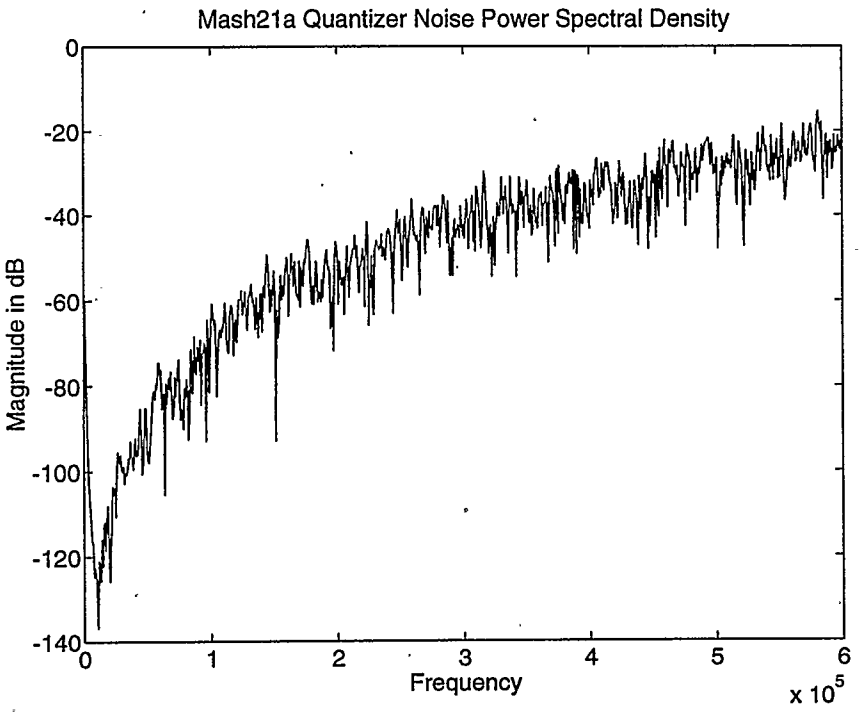


Figure 5.9. The Noise PSD for the Mash21a converter

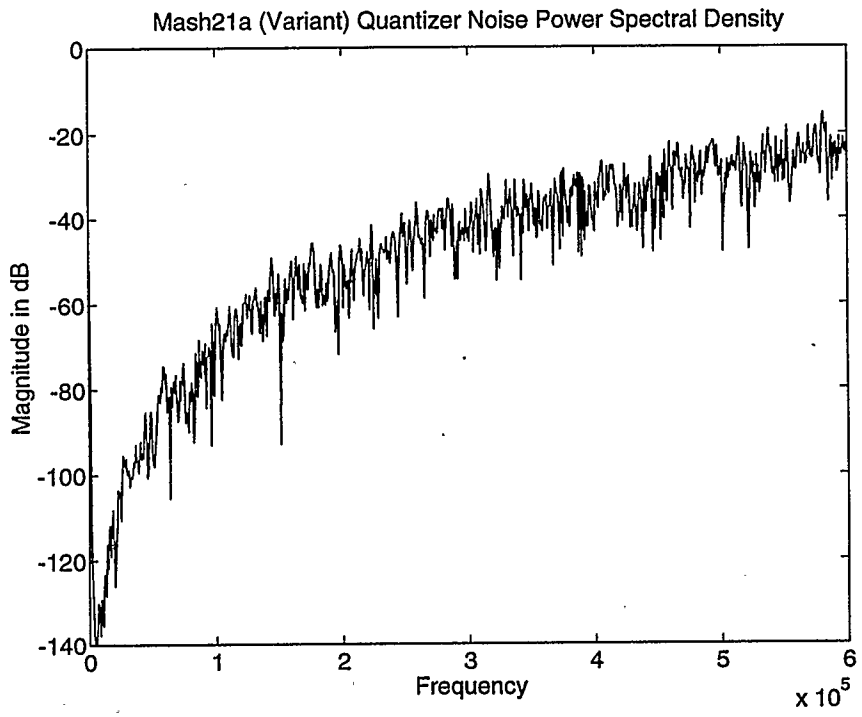


Figure 5.10. The Noise PSD for the modified Mash21a converter

5.5 Characterization of the new converters with regard to signal-to-quantization noise ratio

To evaluate the resolution performance of the new structures, it is necessary to compute the signal-to-quantization noise ratio that is attainable under various converter specifications such as a change in the oversampling ratio or a change in the input signal amplitude (under the no-overload condition). The following results were obtained using the dynamic simulation package SIMULINK in MATLAB where the finite arithmetic effects of the quantizer operation were simulated for two's compliment rounding with saturation.

5.5.1 Signal-to-quantization noise ratio as a function of input signal amplitude

The first sets of data were accumulated for each converter as the input signal amplitude was varied from a low value up to the maximum allowable input signal given by U_1 in (2.35). The input signal contained a simple sinusoid with a designated frequency of $10kHz$ and a fixed oversampling ratio of 64. The performance results of the various new converters may be seen in Figs. 5.11 - 5.13. The simulated results obtained using SIMULINK for the new converters, may be clearly seen to compare favorably with the theoretical results for triple-loop and quadruple-loop converters (using the linear white noise model for the quantizer error).

5.5.2 Signal-to-quantization noise ratio as a function of oversampling ratio

The second set of data values were accumulated for each converter as the oversampling ratio was varied from 40 to 520 while the input signal amplitude was set to the

maximum allowable. The performance results of the various new converters may be seen in Figs. 5.14 - 5.16. Again, the simulated results using SIMULINK may be seen to compare favorably with the theoretical results for triple-loop and quadruple-loop converters (using the linear white noise model for the quantizer error).

5.6 Conclusion

This chapter has presented design techniques for the construction of cascaded $\Sigma - \Delta$ converter configurations. It first discussed the overall design technique in terms of both signal and noise transfer functions, and then presented several new $\Sigma - \Delta$ converters designed using these techniques. It then analyzed these new configurations with regard to signal and noise transfer functions, and finally characterized their performance using their signal-to-quantization noise ratio through simulation. This was performed in two sections, where first the variation in signal-to-quantization noise ratio as a function of input signal amplitude for the new converters was calculated and plotted, and second, the variation in signal-to-quantization noise ratio as a function of oversampling ratio for the new converters was calculated and plotted.

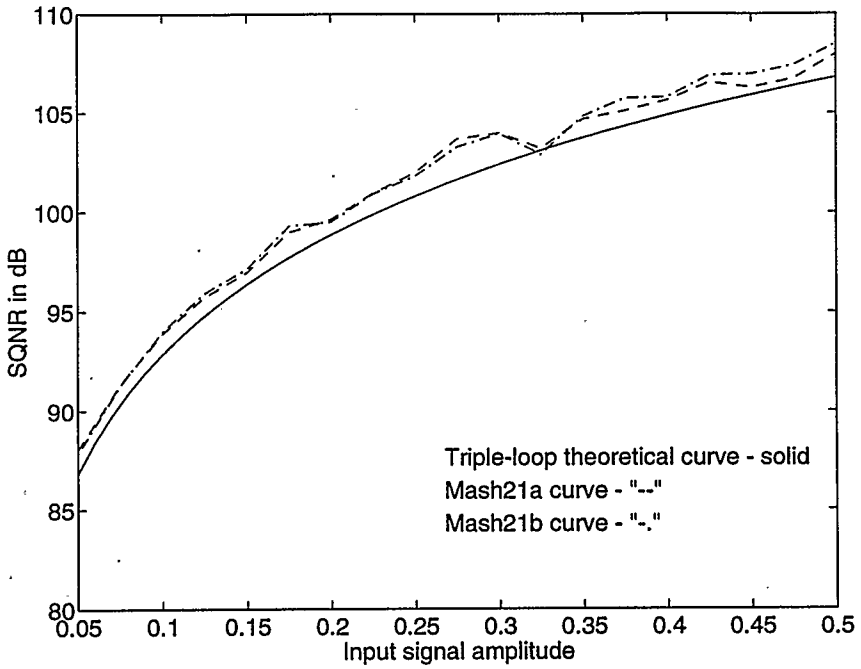


Figure 5.11. SQNR versus input signal amplitude for the Mash21a and Mash21b converters

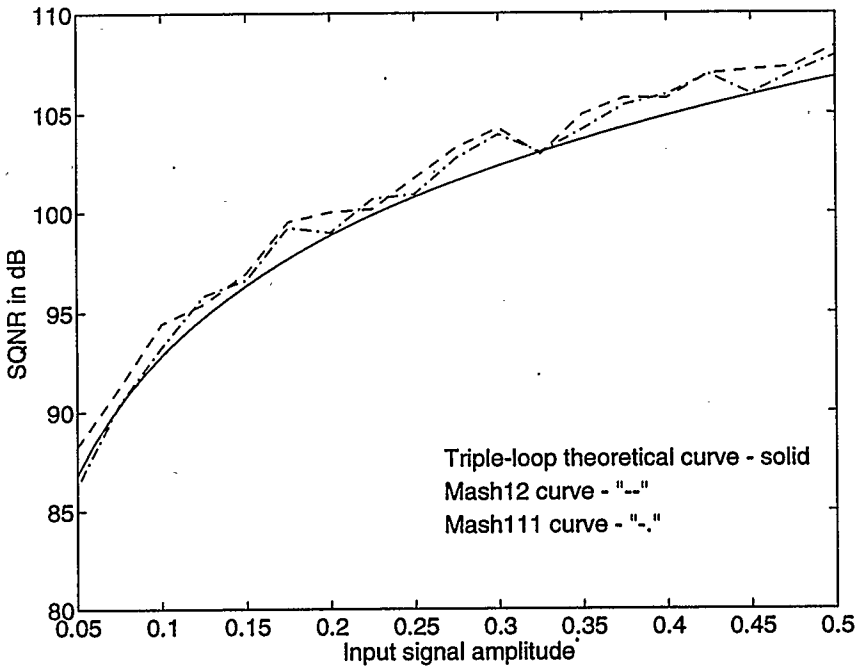


Figure 5.12. SQNR versus input signal amplitude for the Mash12 and Mash111 converters

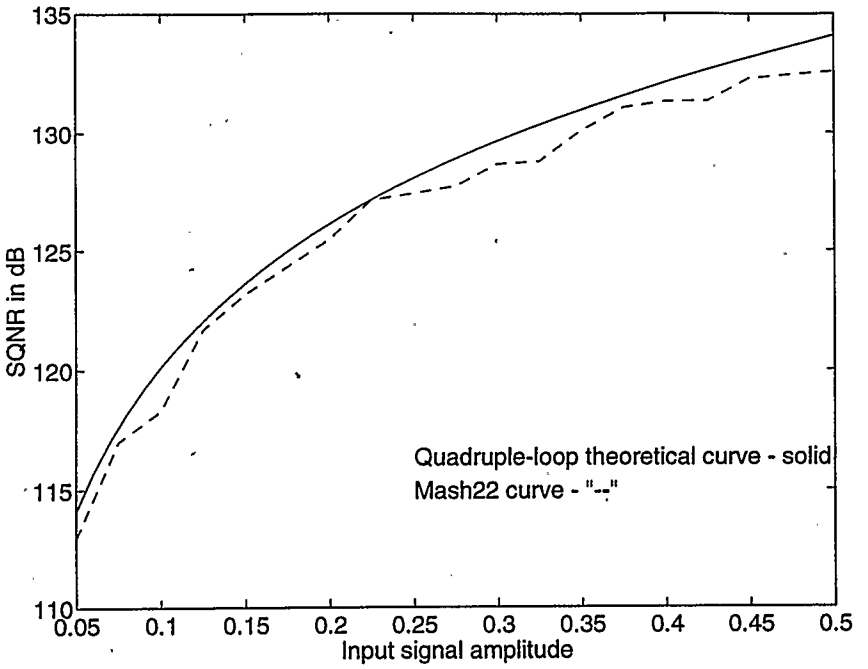


Figure 5.13. SQNR versus input signal amplitude for the Mash22 converter

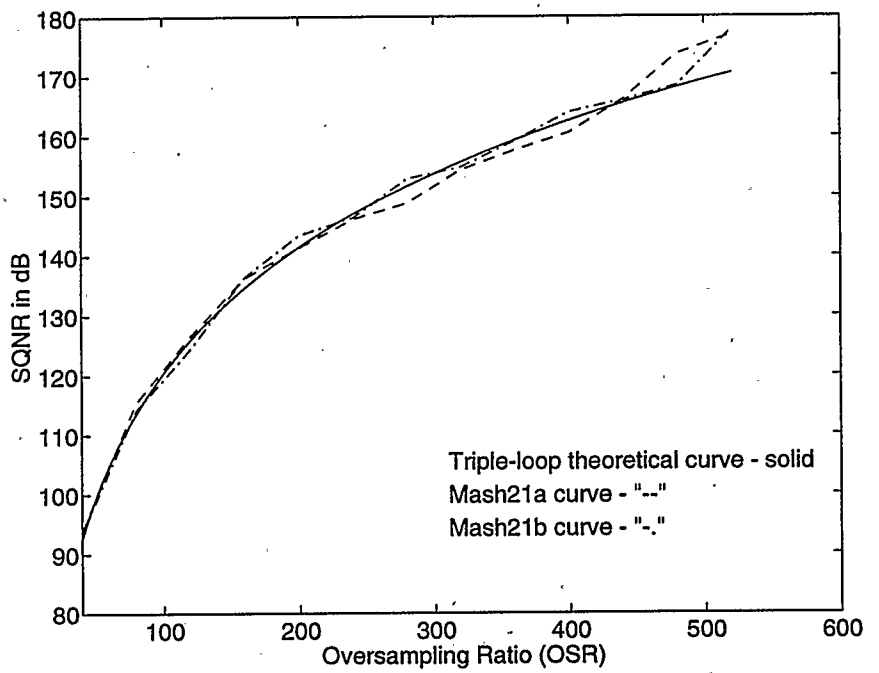


Figure 5.14. SQNR versus OSR for the Mash21a and Mash21b converters

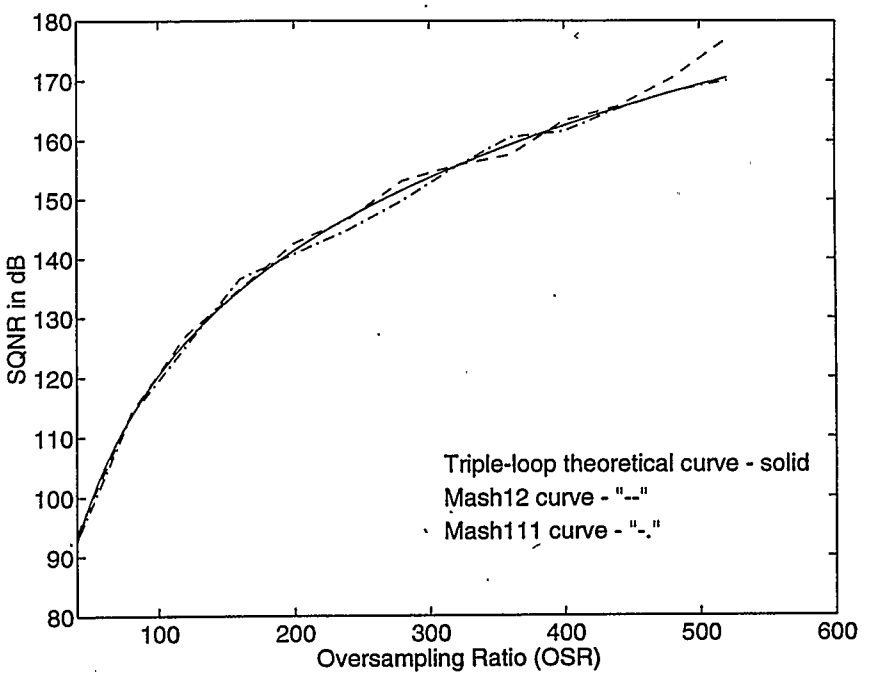


Figure 5.15. SQNR versus OSR for the Mash12 and Mash111 converters

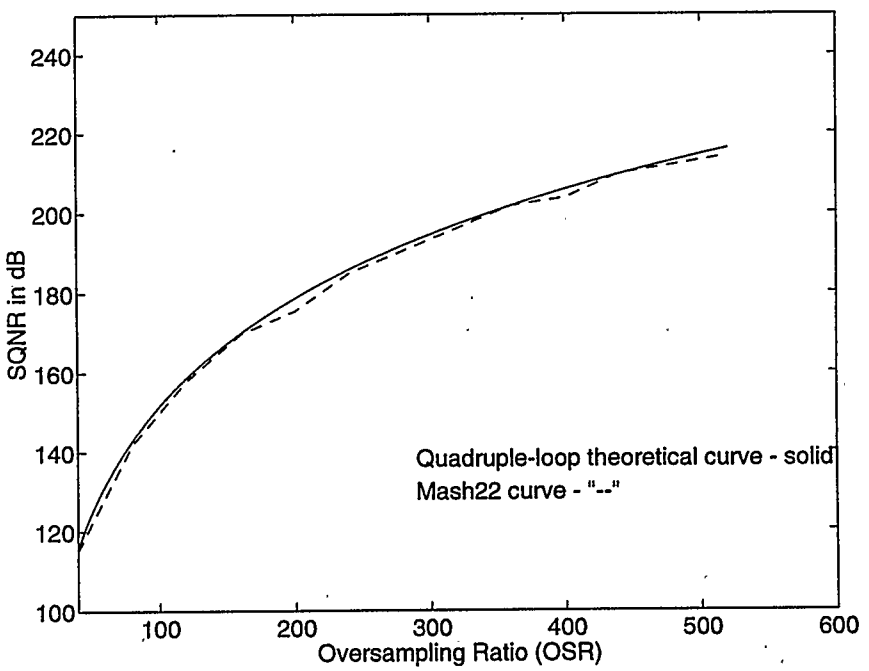


Figure 5.16. SQNR versus OSR for the Mash22 converter

CHAPTER 6

CONCLUSION

6.1 Review of material presented

In Chapter 1 a brief overview of $\Sigma - \Delta$ conversion, with regard first to its original application, then its uses in the field of signal conversion particularly with its applications in modern day interface and communication circuits was given. Also discussed was the basic operation of such converters, and the important issues regarding these converters such as analysis, characterization, and design. Finally, this chapter provided an overview of design methods being utilized to create lowpass $\Sigma - \Delta$ and bandpass $\Sigma - \Delta$ converters.

Chapter 2 then laid the foundation for the next three chapters by first presenting a general state-space representation of a single-quantizer $\Sigma - \Delta$ converter. The state-space equations describing this subsystem were then utilized to derive a closed-form solution for the *granular* quantizer error. Next, an input signal bound to the converter was derived to guarantee that no overloading of the quantizer would occur for the two cases of a) $x(0) = 0$ (the initial states are reset), and b) $x(0) \neq 0$ (the initial states are nonzero). Some interrelationships between various converter parameters were then discussed and finally the utility of the derived theorems was shown with some application examples.

Chapter 3 continued the work started in chapter 2 by deriving a closed-form solution for the quantizer output signal based on the general state-space equations (2.1)

and (2.2) describing the class of single-quantizer $\Sigma - \Delta$ converters. It also derived an open loop equivalent system seen in Fig. 3.1. Application results using these solutions were then compared to results obtained using general difference equations. Finally, discussed were stability issues of $\Sigma - \Delta$ converters with regard to circuit parameters and input signal amplitude. The chapter then concluded with some application examples illustrating these points.

Chapter 4 then dealt with the issue of spectral analysis of $\Sigma - \Delta$ converters and how such spectral information is used to analyze converter behavior and determine converter performance. It examined some of the different techniques used by designers to determine spectral behavior of $\Sigma - \Delta$ converters with a particular look at the three spectral analysis methods of

a) The linear white noise model of the quantizer error.

b) The Fourier series representation of the error sequence.

c) First and second moments of the error sequence derived using characteristic functions.

Some applications examples were then presented and finally all discussed methods were used in the calculation of the converter signal-to-quantization noise ratio.

Chapter 5 concluded the research material presented by presenting design techniques for the construction of cascaded $\Sigma - \Delta$ converters applicable to either A/D or D/A converter circuits. It first formulated the overall design technique in terms of the signal and noise transfer functions, and then presented various new $\Sigma - \Delta$ converter configurations designed using these techniques. The third section of this chapter then analysed the new converter configurations with regard to their signal

and noise transfer functions and finally, the operation of these new converters was characterized with regard to their signal-to-quantization noise ratio through simulation. This was examined in two subsections, the first determining the variation in signal-to-quantization noise ratio as a function of input signal amplitude and the second determining signal-to-quantization noise ratio as a function of the oversampling ratio.

6.2 Proposed areas for future research and improvements

This thesis has presented several formal analytical methods that will give designers of $\Sigma - \Delta$ converters insight into system behavior which cannot be obtained with the application of *ad hoc.* design techniques and simulations alone. It may be said that while simulations are useful tools for analysing final designs, they do not provide sufficient knowledge into certain system behaviour that may occur only for particular given input signals. Such shortcomings of simulation tools (such as the use of a *independent identically distributed* noise model for the quantizer) are exactly why formal analysis methods are so valuable.

The value of closed-form solutions for determining granular error and quantizer output and their subsequent use in spectral analysis has been clearly due to the reduction in the number of computations required to obtain a solution, but the closed-form solutions obtained in this thesis have also been shown to operate for only a certain class of $\Sigma - \Delta$ converters. This class may include a large number of multiloop and cascade $\Sigma - \Delta$ converters, but it is still limited to those configurations that contain integer valued multipliers only. This limitation may be clearly traced to the two relations used in deriving (2.12) and (3.6), namely

$$\langle r_1 + K\langle r_2 \rangle \rangle = \langle r_1 + K r_2 \rangle$$

and

$$\lfloor r + K\lfloor s \rfloor \rfloor = \lfloor r \rfloor + K\lfloor s \rfloor$$

which dictate that K must be an integer and therefore the matrix products $C(A + B_2C)^lB_2$ and CA^lB_2 are required to be integers for each $l = 0, 1, \ldots, \infty$ (implying that all multipliers in the subsystem $\mathcal{N}$ given in Fig. 2.1 must be integer valued multipliers). New analytical methods must be examined and developed that would make such general state-space equations applicable to $\Sigma - \Delta$ converter structures that contain noninteger valued multipliers as well, such as the TOSLAP converter of Fig. 1.11. This would permit designers to analyse arbitrary $\Sigma - \Delta$ converter structures using only general equations and thus preclude the necessity of writing and solving a set of difference equations for each different converter configuration.

A second area that needs to be explored in detail, is that of the analysis of $\Sigma - \Delta$ converters under quantizer overload, for at present there are not any analytical tools that this author is aware of that could determine the quantizer error when the signal input to the quantizer exceeds the no-overload or granular range.

Finally, an area that is often overlooked and quite naturally needs more development is that of formal design methods to produce $\Sigma - \Delta$ converters for given specifications (either lowpass or bandpass applications). These methods must fully exploit the complimentary nature of such structures unlike the trial and error development method with optimization packages that usually occurs today.

6.3 Concluding remarks

It may be said with all honesty, that result of this research is built upon the foundation of such pioneering analytical work in the field of $\Sigma - \Delta$ converters by such researchers as Candy, Temes, Gray, and others. To paraphrase my old highschool

football coach, " We have taken the ball passed to us by others, and we have run further down the field with it.", signifying that while some work has been accomplished, there is much more work yet to be done that will be accomplished by others. However, as one whom is thoroughly captivated by oversampled $\Sigma - \Delta$ converters, this field will always be of constant interest to me, due to the continual introduction of new and innovative techniques and approaches of analysis.

REFERENCES

[Ben48] W. R. Bennett. Spectra of quantized signals. *Bell Syst. Tech. J.*, pages 446–472, 1948.

[BF77] C. W. Barnes and A. T. Fam. Minimum norm recursive digital filters that are free of overflow limit cycles. *Trans. Circuits Syst.*, CAS-24:569–574, Oct. 1977.

[BF94] R. T. Baird and T. S. Friez. Stability analysis of high-order delta-sigma modulation for adc's. *Trans. Circuits Syst.*, CAS-41:59–62, Jan. 1994.

[BN94] T. P. Borsodi and B. Nowrouzian. A survey of oversampled sigma-delta a/d and d/a converters for digital audio applications. In *Proceedings of the 37th Midwest Symposium on Circuits and Systems, Lafayette, Louisiana, 1994*, pages 1208–1211. IEEE Circuits and Systems Society, 1994.

[BN95] T. P. Borsodi and B. Nowrouzian. Closed-form solution of granular quantization error for a class of sigma-delta modulators. In *1995 IEEE International Symposium on Circuits and Systems, Seattle, Washington, 1995*, pages 629–632. IEEE Circuits and Systems Society, 1995.

[Can74] J. C. Candy. A use if limit cycles oscillations to obtain robust analog-to-digital converters. *IEEE Trans. Commun.*, COM-22:298–305, Mar. 1974.

[Can85] J. C. Candy. A use of double integration in sigma-delta modulation. *IEEE Trans. Commun.*, COM-33:249–258, Mar. 1985.

[CB81] J. C. Candy and O. J. Benjamin. The structure of quantization noise from sigma-delta modulation. *IEEE Trans. Commun.*, COM-29:1316–1323, Sept. 1981.

[CNLS90] K. C. H. Chao, S. Naddeem, W. L. Lee, and C. G. Sodini. A higher order topology for interpolative modulators for oversampling a/d conversion. *IEEE Trans. Circuits Sys.*, CAS-37:309–318, Mar. 1990.

[CPG47] A. G. Clavier, P. F. Panter, and D. D. Grieg. Distortion in a pulse count modulation system. *AIEE Trans.*, 66:989–1005, 1947.

[CT92] J. C. Candy and G. C. Temes. Oversampling methods for a/d and d/a conversion. In J. C. Candy and G. C. Temes, editors, *Oversampling Delta-Sigma Data Converters*, pages 1–25. IEEE Press, 1992.

[DR58] W. B. Davenport and W. L. Root. *An Introduction to the Theory of Random Signals and Noise.* McGraw-Hill, 1958.

[GCW89] R. M. Gray, W. Chou, and P. W. Wong. Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs. *IEEE Trans. Commun.*, COM-37:956–967, Sept. 1989.

[Ge89] Gailus and *et al.* Method and arrangement for a sigma-delta converter for bandpass signals. United States Patent, Aug. 1989. Patent No. 4,857,928.

[Gra89] R. M. Gray. Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input. *IEEE Trans. Commun.*, COM-37:588–599, June 1989.

[Gra90] R. M. Gray. Quantization noise spectra. *IEEE Trans. Information Theory*, IT-36:1220–1244, 1990.

[HJ93] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1993.

[HKB92] N. He, F. Kuhlmann, and A. Buzo. Multiloop sigma-delta quantization. *IEEE Trans. Information Theory*, IT-38:1015–1028, May 1992.

[Iwe69] J. E. Iwersen. Calculated quantizing noise of single-integration delta-modulation coders. *Bell Syst. Tech. J.*, pages 2359–2389, Sept. 1969.

[IYM62] H. Inose, Y. Yasuda, and J. Murakami. A telemetering system using code modulation - delta-sigma modulation. *IRE trans. Space Elect. Telemetry*, SET-8:204–209, Sept. 1962.

[JSF93] S. Jantzi, W. M. Snelgrove, and P. Ferguson. A fourth-order bandpass sigma-delta modulator. *IEEE J. Solid-State Circuits*, SSC-28:282–291, March 1993.

[JSS91] S. Jantzi, R. Schrier, and M. Snelgrove. Bandpass sigma-delta analog-to-digital conversion. *IEEE Trans. Circuits Sys.*, CAS-38:1406–1409, Nov. 1991.

[LC88] L. Longo and M. A. Copeland. A 13-bit isdn-band adc using two-stage third order noise shaping. In *Proceedings of the 1988 Custom Integrated Circuit Conference, 1988*, pages 21.2.1–4, June 1988.

[Lju87] L. Ljung. *System Identification.* Prentise-Hall, 1987.

[LT93] L. E. Larson and G. C. Temes. Signal conditioning and interface circuits. In S. K. Mitra and J. F. Kaiser, editors, *Handbook for Digital Signal Processing*, pages 677–718. Wiley-Interscience, 1993.

[Rib91] D. B. Ribner. A comparison of modulator networks for high-order oversampled sigma delta analog-to-digital converters. *IEEE Trans. Circuits Sys.*, CAS-38:145–159, Feb. 1991.

[Ric54] S. O. Rice. Mathematical analysis of random noise. In N. Wax, editor, *Selected Papers on Noise and Stochastic Processes*, pages 133–294. Dover, 1954.

[RL94] S. Rangan and B. Leung. Quantization noise spectrum of double-loop sigma-delta converter with sinusiodal input. *IEEE Trans. Circuits Sys.*, CAS-41:168–173, Feb. 1994.

[SS89] R. Schreier and M. Snelgrove. Bandpass sigma-delta modulation. *Electron. Letters*, 25:1560–1561, Nov. 1989.

[TPH91] A. M. Thurston, T. H. Pearce, and M. J. Hawksford. Bandpass implementation of the sigma-delta a-d conversion technique. In *Proceeding of the IEE International Conference on Analogue-to-Digital and Digital-to-Analogue Conversion, Swansea, England*, pages 81–86, Sept. 1991.

[UHKI88] 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*, ASSP-36:1899–1905, Dec. 1988.

[WG90] P. W. Wong and R. M. Gray. Two-stage sigma-delta modulation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-38:1937–1952, Nov. 1990.

[ZD63] L. A. Zadeh and C. A. Desoer. *Linear System Theory: The State Space Approach.* Robert. E. Krieger Publishing Company, 1963.