REDUCED-COMPLEXITY DIGITAL SINUSOID GENERATORS AND OVERSAMPLED DATA CONVERTERS

Thesis by
Michael J. Flanagan

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1995
(Defended June 13, 1994 and June 15, 1994)

ACKNOWLEDGEMENTS

This thesis is dedicated to the memory of Prof. Edward C. Posner (1933-1993). Ed was my first thesis advisor and is a sorely missed friend. I thank God for all good things; especially for the opportunity to have known Ed and for the three years I was able to work with him. I would like to thank Prof. Robert McEliece for becoming my official thesis advisor after Ed's untimely death. Prof. McEliece's advice and encouragement are greatly appreciated. I would also like to thank my defense committee: Professors McEliece, Vaidyanathan and Galton and Drs. Zimmerman and Satorius.

Dr. George Zimmerman has been an important teacher to me. George has supervised almost all of the research and design I have done in the past four years, both at Caltech and the Jet Propulsion Laboratory (JPL). George has always been my "unofficial thesis advisor," and has performed this task above and beyond the call of duty. I am grateful for his keen insight, his kind mentoring and his friendship.

I cannot imagine how different these past four years would have been without my academic part-time employment at JPL. The JPL projects that I have worked on have shaped and motivated this thesis in a significant manner. My friendship with members of Helmut Wilck's Digital Projects Group is gratefully acknowledged. I would also like to thank the National Science Foundation and JPL for financial support over the last four years.

I am nothing without my family. The love of my dear wife, Jeannine, has been a constant source of encouragement and strength. My beloved son, James Walter, has brought more happiness to my life than I thought was possible. I thank my Mother and Father for their love and prayers and for being my earliest and most important teachers. The support and love of my brothers, Phil, Peter and Jack, are greatly appreciated. I am especially thankful for all of the help Peter has given me during my stay in California. The love, prayers and support of my godmother, Kathleen, are also lovingly acknowledged. These acknowledgements would be incomplete without thanking Michael Messier for being an excellent friend. Finally, a long-distance "hi-dilly-ho" and thanks go out to my distant friends: Jayquatta, Doug, Laura and Ted.

ABSTRACT

This thesis separately addresses two important issues in signal processing: digital sinusoid generators and oversampled data converters. The first part of the thesis addresses noise additive, or dithering, techniques that exponentially reduce the complexity of digital sinusoid generators for a given level of spur performance. With the appropriate dither signals the quantization noise can be rendered nearly white and free of large spurs, or periodic error components, without recourse to large look-up tables. New analysis shows that when the phase dither signal is the sum of M uniform white variates, the phase spurs are at a level of $-6(M+1)$ dBc per look-up phase bit instead of the usual -6 dBc per phase bit in a non-dithered system. This exponentially reduces the complexity of the digital sinusoid generator for a given spur requirement at the expense of linearly increasing the nearly-white quantization noise.

The second part of the thesis presents two metric-based approaches to the design of oversampled data converters (ODCs). The first approach leads to an architecture which is derived based on the minimization of a causal, constrained-memory, power-spectral-distortion metric. This architecture is compared to standard $\Delta\Sigma$ modulators and shown to have superior noise performance under some conditions.

Another metric-based approach to the design of ODCs uses a more general distortion metric and incorporates elements of vector quantization, eigensystems and analysis of the discrete prolate spheroidal wave functions. This enables the application of vector quantization theory to oversampled data converters. A vector-quantizer-based ODC architecture called the eigenmodulator is motivated and analyzed. Rate-distortion results are presented for the important case of a band-limited Gaussian input. When both the complexity of the eigenmodulator and the oversampling ratio become large, it is shown that the distribution of the output vectors in an important transform space becomes joint Gaussian. This is shown to be important in light of the centroid condition for an optimal vector quantizer. The implication of this result on the choice of output scaling for the single-bit data converter is addressed.

Table of Contents

Chapter 1

Introduction and Overview

This thesis addresses two important issues in signal processing: the generation of digital sinusoids and oversampled data conversion. Digital sinusoid synthesizers (also commonly referred to as Numerically-Controlled Oscillators, or NCOs) are digital machines that generate a discrete-time signal that is an approximation to an ideal discrete-time sinusoid [1]. The approximation is generally inexact because of quantization effects caused by finite-word length amplitude and phase representations. Oversampled data conversion refers to the digital-to-analog or analog-to-digital conversion of a signal whose analog bandwidth is smaller than the digital clock rate at which it is sampled [2]. An oversampled data converter typically has few quantization levels at the stage where the conversion occurs and is followed by a linear filter designed to reject much of the error in the coarse quantization.

The obvious connection between digital sinusoid synthesis and oversampled data conversion is exemplified by a system in which the highly oversampled output of a digital sinusoid generator is sent to a coarse-resolution digital-to-analog converter (DAC) followed by a low-pass filter to create an analog waveform. The output of the digital sinusoid generator could be highly oversampled in the sense that it only generates frequencies from DC to a small fraction of the digital clock rate. The motivation for this type of a system will be considered in light of DAC non-linearities after a preliminary discussion of digital sinusoid generators and oversampled data converters. This chapter concludes with a presentation of the major

results of the thesis.

1.1 Digital Sinusoid Generators

Digital sinusoid synthesizers are being employed more and more in receivers and transmitters as the trend to replace analog systems with inexpensive and reliable digital systems continues. These digital sinusoid synthesizers are frequently used in mixing applications [3]. In digital receivers, the desired signal is digitized while still at an intermediate frequency (IF) and then digitally multiplied by a digital sinusoid to translate the signal to a desired frequency range. In a direct-digital frequency synthesizer (DDS), the output of the digital sinusoid generator is sent to a digital-to-analog converter and subsequently can be used in a variety of analog signal processing systems.

The digital sinusoid generator architecture considered throughout the thesis is based on the synchronous discrete-time, discrete-amplitude design of Tierney et. al. [1]. Fig. 1 illustrates the basic building blocks: a phase accumulator, a phase truncator, a phase-to-amplitude converter and an amplitude truncator. The phase accumulator outputs the sum of the previous phase and the phase increment. Therefore, the output of the phase accumulator is the phase of the desired sinusoid scaled by 2π. In this architecture, both the phase and the phase increment are measured in cycles, not radians. The phase increment specifies the output frequency. Let the phase increment be represented using B bits and view this as an unsigned fractional quantity, i.e., when the phase increment bits are 01000... interpret them to be $(0.01000...)_2 = 0.25$. The output frequency of the digital sinusoid generator will be the digital clock rate times the fractional phase increment. Increasing the number of bits used to represent the phase increment increases the frequency resolution at the expense of increasing the complexity of the phase accumulator since the phase is represented using the same number of bits as the phase increment. Since the phase (expressed in radians) of the sinusoid is only important modulo 2π, the output of the phase accumulator is permitted to overflow. In fact, the regular overflow of the phase accumulator establishes the periodicity

of the output signal.

In order to divorce frequency resolution requirements from system complexity as much as possible, the number of phase bits used to specify an output amplitude is generally less than the number of bits required to satisfy a particular frequency resolution [4]. This reduction in the number of bits is referred to as quantization since it is a non-linear, many-to-one mapping into discrete quantum intervals. While quantization is almost always considered in the classical literature as the mapping of a continuous amplitude onto a discrete set, it will be useful for the remainder of this thesis to consider digital word-length reduction as a form of quantization. This is because digital word-length reduction is the mapping of one discrete set onto a smaller discrete set. The error associated with this many-to-one mapping is called the *quantization error*. The quantization operation can always be conveniently modeled by the addition of a properly defined error signal.

One way to implement the phase-to-amplitude converter is by using a memory look-up table. This approach will be considered throughout the rest of the thesis because of its ubiquity in the literature and in commercial systems. The size of the memory look-up table grows exponentially in the number of address bits and linearly with the length of the words contained in the table. Because of this exponential dependence, the complexity of the digital sinusoid generators considered in the remainder of this thesis will be largely driven by the number of phase bits presented to the look-up table. While there are other ways to implement the phase-to-amplitude conversion, the complexity of all approaches generally will increase as the number of input and output bits increases.

The amplitude truncator reduces the word length of the output digital signal to interface it with other systems. For example, the number of output bits required may be driven by a digital-to-analog converter or by the input to a digital multiplier. While there is frequently no need for a physical embodiment of an amplitude truncator in the conventional architecture (since the look-up table would only generate as many bits as was needed), it is included symbolically to be consistent with noise addition schemes that will be presented later.

The quantization errors introduced by phase and amplitude truncation are periodic be-

cause they are generated when periodic signals pass through a memoryless non-linearity. Phase quantization errors are created by the word-length reduction of the phase out of the phase accumulator going into the look-up table. The phase accumulator output is a sawtooth ramp with a period of at most 2^B clock cycles, where B is the number of bits out of the phase accumulator. The phase quantization is memoryless because the phase quantizer makes the same error every time the same input is presented. Amplitude quantization errors are created by the process of representing a sinusoidal sample that generally requires an infinite number of bits with a finite number of bits. Since the quantization errors are periodic the error spectrum is littered with discrete frequency components that are commonly referred to as *spurs*. In general, when quantizing a signal to b bits, the resulting quantization error is $O(2^{-b})$ in amplitude. As a result, the spur power magnitudes are $O(2^{-2b})$ or approximately 6 dBc per bit down from the power in the fundamental frequency component.

In this conventional digital sinusoid generator with a look-up table, in order to improve the spurious performance it is necessary to increase the number of phase look-up bits and/or the number of amplitude bits. The former results in an exponential increase in system complexity and the latter results in a linear increase in system complexity and requires the use of higher resolution digital-to-analog converters and multipliers, which may not be feasible.

1.2 Oversampled Data Converters

Oversampled data converters have become increasingly popular in the last decade due to advances in digital VLSI processing. Oversampled data converters shift much of the signal processing burden from the analog to the digital domain. As a result, the use of less complex or precise analog components can be compensated by an increase in the complexity of the digital signal processing system. The latter is typically more acceptable in modern commercial systems because digital circuitry tends to be more robust and insensitive to circuit imperfections than analog circuitry [2].

Data conversion refers to the operations of converting a digital waveform into an analog waveform (digital-to-analog conversion, or D/A conversion) and converting an analog waveform into a digital waveform (analog-to-digital conversion, or A/D conversion.) On a theoretical level described in greater detail later, this work treats both A/D and D/A converters without distinction as they both involve the mapping of a sequence with a continuous (or discretely valued with high resolution) amplitude distribution to and from a sequence with a discrete amplitude distribution. The exact mapping is often based on a fidelity criterion such as the mean-squared error.

Oversampled data converters (ODCs) may be characterized by two important traits. As their name implies, oversampled data converters operate on signals that are either originally (D/A) or finally (A/D) sampled in the digital domain above their Nyquist rate. The *oversampling ratio*, R, is defined to be the ratio of the sample rate to the Nyquist rate. The second important trait is that oversampled data converters use quantizers of coarser resolution than the equivalent Nyquist-rate quantizers. In the limiting (and popular) case, single-bit quantizers are employed, which results in relatively simple analog circuitry. In an A/D converter, a single-bit quantizer uses a single voltage-threshold comparator. In a D/A converter, a single-bit quantizer uses a single current source. As will be discussed shortly, the use of coarse-resolution quantizers can also result in a more linear system. This is of critical importance when generating sinusoids with high spectral purity. However, as the number of bits in the quantizer decreases, the amount of quantization noise increases, as discussed in the previous section. This is the motivation for oversampling. By restricting the support of the desired signal in the frequency domain, one can heuristically place the quantization noise predominantly into frequency regions unoccupied by the desired signal. Then ideal frequency-selective filtering could remove much of the quantization distortion. It is worth commenting that in the case where the desired signal is a sinusoid it is true that the frequency support is zero. However, this thesis considers the generation of sinusoids whose frequencies span a given range: 0 to 10 MHz, for example. In this example, the bandwidth is said to be 10 MHz and instead of using an analog bandpass filter with arbitrarily small

support and a programmable center frequency (which would involve relatively complicated analog circuitry) an analog low-pass filter with a DC to 10 MHz passband would follow the D/A converter.

We now consider the linearity of realizable quantizers. In practice, the analog levels of a quantizer are not going to be perfectly spaced as in an ideal quantizer. Therefore, any realizable quantizer can be expressed statically as an ideal quantizer followed by a polynomial of degree $2^b - 1$ where b is the number of bits in the quantizer. Assume that we are able to replace the quantizer with an additive error source and that the spectrum of the additive error contained no discrete frequency components. Assume also that our input signal is a sinusoid. It will be shown in Chapter 2 that this assumption can be verified under appropriate circumstances. Then the polynomial will create 2nd, 3rd, 4th and higher harmonics of the input signal which corrupt the output spectrum as spurs. One way to minimize the power in these spurs is to control the magnitudes of the coefficients in the quantizer non-linearity polynomial. This is often not possible since the coefficient magnitudes are driven by the number of bits of resolution. Another possible solution is to use a single-bit data converter. In this case, $b = 1$, and the DAC "non-linearity" polynomial is a first-degree polynomial which can be contrived to contribute no spurs.

1.3 Thesis Results

This thesis is divided into two parts. Chapter 2 is devoted to the presentation and analysis of practical dithering techniques that reduce the complexity of digital sinusoid generators. Chapters 3, 4 and 5 study problems in oversampled data conversion. Chapter 6 details future work.

The central feature of Chapter 2 is the insertion of an additive noise signal, or *dither signal*, prior to the word-length reductions in both the amplitude and phase in a digital sinusoid generator. The dither signal added prior to the phase quantization is referred to as the phase dither and the dither signal added prior to the amplitude quantization is referred to

as the amplitude dither. It is shown that with the appropriate dither signals, the quantization noise can be rendered nearly white and free of large spurs without recourse to large look-up tables. Specifically, the addition of uniformly distributed white dither prior to amplitude quantization eliminates the generation of amplitude spurs. The elimination of amplitude quantization spurs is achieved at the expense of increased noise which is approximately white. This is tolerable in a large class of systems that have large bandwidths or process signals with all but the largest signal-to-noise ratios. In contrast to amplitude dithering, it is shown that systems with finite complexity cannot eliminate the spurs due to phase quantization using dither signals composed of sums of uniformly distributed white variates. Instead, when the phase dither signal is the sum of M uniform white variates, the phase spurs are at a level of $-6(M+1)$ dBc (deciBels with respect to the carrier) per look-up phase bit instead of the usual -6 dBc per phase bit in the conventional, non-dithering ($M = 0$) case. This is the key result of the first part of the thesis. The use of this kind of a phase dither signal permits the use of fewer look-up phase bits to maintain a given spur performance level and results in an *exponential* reduction in the complexity of the digital sinusoid generator. Chapter 2 also addresses the use of long Linear Feedback Shift Registers (LFSRs) to generate an efficient approximation to the uniform variates subject to given spurious specifications.

The second part of the thesis studies oversampled data converters. The central feature is the use of distortion metrics in the design of oversampled data converters. The key result is the connection between oversampled data converters and vector quantization [6]. The second part of the thesis lays the groundwork for future oversampled data converters that employ vector quantizers.

Chapter 3 defines an Oversampled Data Converter Problem used in the subsequent chapters and presents, analyzes and simulates a synthetic approach to the design of oversampled data converters. An architecture is derived based on the minimization of a causal, constrained-memory, power-spectral-distortion metric and is called the Synthesis architecture. The Synthesis architecture is shown to be related to a popular class of oversampled data converters called $\Delta\Sigma$ modulators. Under common quantization noise assumptions, the

Synthesis architecture is shown to theoretically have superior noise performance over the first-order $\Delta\Sigma$ modulator for all oversampling ratios and superior noise performance over all $\Delta\Sigma$ modulators of *arbitrary* order when the oversampling ratio is less than 2.862. Subsequent simulations test the accuracy of the theoretical predictions and investigate the requisite complexity of the Synthesis architecture in order to approximate asymptotic (infinite complexity) noise performance as the oversampling ratio becomes small.

Chapter 4 presents a novel vector quantization approach to the analysis and design of oversampled data converters that incorporates elements of eigenanalysis and the discrete prolate spheroidal wave function analysis of Slepian [5]. It is shown that oversampled data converters can be analyzed and designed in the context of vector quantizers. The use of an appropriate transform space permits rate-distortion analysis for the important case of a band-limited Gaussian input. This provides bounds for the minimum distortion achievable in an oversampled data converter as a function of oversampling ratio and complexity. Results for the special case of asymptotically large complexity are presented. The transform space analysis motivates a new oversampled data converter architecture called the eigenmodulator which employs a vector quantizer.

Chapter 5 analyzes the eigenmodulator in the context of vector quantization theory. This lays the groundwork for future applications of vector quantization theory to the analysis and design of oversampled data converters. The special case of the single-bit quantizer is considered exclusively. Under certain asymptotic conditions where the oversampling ratio, in addition to the block size, becomes large, it is shown that the mapping of the binary output vectors into an important transform space has a joint Gaussian distribution. This fact combined with the known conditions for an optimal vector quantizer [6] motivates an analysis of the eigenmodulator when the input is joint Gaussian with variance σ^2 and oversampled by the oversampling ratio, R. An application of vector quantization theory leads to the conjecture that a nearly optimal choice of the $\pm c$ output levels in the single-bit data converter is $c = \sigma R^{\frac{1}{2}}$, as the oversampling ratio, R, becomes large.

Chapter 6 presents items for further research based on the results of the previous chapters.

References

1. J. Tierney, C. Rader, B. Gold, "A Digital Frequency Synthesizer," *IEEE Transactions on Audio and Electroacoustics,* vol. AU-19, No. 1, pp. 48-57,March 1971.

2. *Oversampled Data Converters: Theory, Design and Simulation,* edited by J. Candy, G. Temes, IEEE Press, New York, 1992.

3. S. Hinedi, "NASA's Next Generation All-Digital Deep Space Network Breadboard Receiver," *IEEE Transactions on Communications*, vol. COM-41, pp. 246-257, January 1993.

4. R. J. Zavrel, Jr. and G. Edwards, The DDS Handbook: Second Edition, Stanford Telecommunications, Santa Clara, CA, 1990.

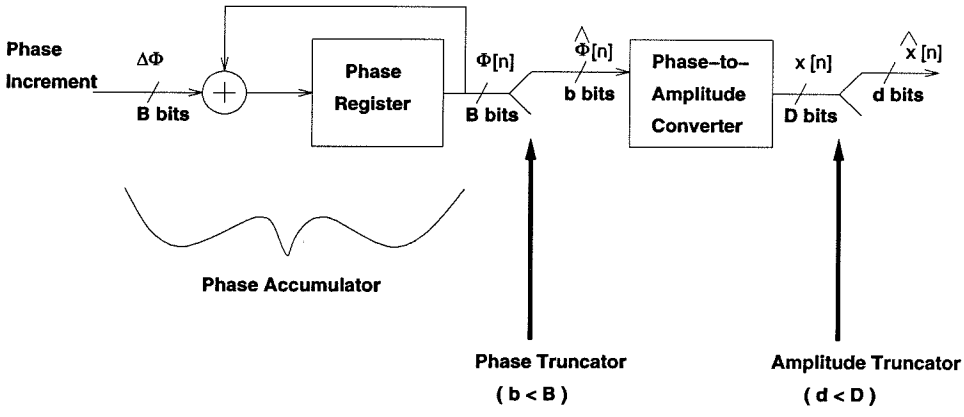
5. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty - V: The Discrete Case," *Bell Systems Technical Journal,* vol. 57, pp. 1371-1430, May-June 1978.

6. A. Gersho and R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, 1991.

Figures

1. A digital sinusoid generator architecture

Figure 1: A Digital Sinusoid Generator Architecture



Chapter 2

A Dithered Digital Sinusoid Synthesizer with Reduced Complexity

2.1 Introduction

This chapter presents and analyzes a technique for reducing the complexity of a digital sinusoid synthesizer for a given level of spurious performance. This reduction is accomplished through dithering both amplitude and phase values prior to word-length reduction. The analytical approach developed for analog quantization is used to produce new bounds on spur performance in these dithered systems. Amplitude dithering allows output word-length reduction without introducing additional spurs. The effects of periodic dither similar to that produced by a linear feedback shift register (LFSR), or pseudo-noise (PN) generator, are analyzed. A phase dithering method provides a spur reduction of $6(M+1)$ dB per phase bit when the dither is the sum of M independent, uniform variates. While the reduction in complexity is made at the expense of an increase in system noise, the noise can be made approximately white, making the noise power spectral density small. This dithering technique permits the use of a smaller number of phase bits addressing sinusoid look-up tables,

resulting in an exponential decrease in system complexity. Amplitude dithering allows the use of less complicated multipliers and narrower data paths in purely digital applications, as well as the use of coarse-resolution, highly-linear digital-to-analog converters (DACs) to obtain spur performance limited by the DAC linearity rather than its resolution.

It is well-known that adding a dither signal to a desired signal prior to quantization can render the quantizer error independent of the desired signal [1, 2, 3]. Classic examples of this work deal with the quantization of analog signals. Advances in digital signal processing speed and large scale integration have led to the development of all-digital receiver systems, direct digital frequency synthesizers and direct digital arbitrary waveform synthesizers. In all these applications, since finite word-length effects are a major factor in system complexity, they may ultimately determine whether it is efficient to digitally implement a system with a particular set of specifications. Earlier work [4] has presented a technique for reducing the complexity of digital oscillators through phase dithering with the claim of increased frequency resolution. Recent research [5] has suggested mitigation of finite-word-length effects in the synthesis of oversampled sinusoids through noise shaping. This chapter shows how the analysis techniques used for quantization of analog signals can be applied to overcome finite-word-length effects in digital systems. The analysis shows how appropriate dither signals can be used to reduce word lengths in digital sinusoid synthesis without suffering the normal penalties in spurious signal performance. Furthermore, the dithering technique presented is not limited to the synthesis of oversampled signals.

Conventional methods of digital sinusoid generation [6], e.g., Fig. 1, result in spurious harmonics (spurs) which are caused by finite word-length representations of both amplitude and phase samples [7]. Because both the phase and amplitude samples are periodic sequences, their finite word-length representations contain periodic error sequences, which cause spurs. The spur signal levels are approximately 6 dB per bit of representation below the desired sinusoidal signal.

The dithering technique presented in this chapter reduces the representation word length without increasing spur magnitudes by first adding a low-level random noise, or dither, signal

to the amplitude and/or the phase samples, which are originally expressed in a longer word length. The resulting sum, a dithered phase or amplitude value, is truncated or rounded to the smaller, desired word length. Of course, either the amplitude or the phase or both can be dithered. In phase dithering the spurious response is determined by the type of dithering signal employed. In amplitude dithering the spurious response is determined by the original, longer word length. While the amplitude-related spurious is generally related to the phase-related spurious, we will make the pre-dither amplitude word length long enough to satisfy spur power specifications. Then the exact relationship is unimportant, and since the phase dither signal is chosen to be independent of the amplitude dither signal, the amplitude and phase dithering processes can be treated independently.

The next section describes the quantizer model. Amplitude and phase quantization effects are reviewed in Sections 2.3 and 2.4, and simple new bounds on spurious performance in conventional systems are presented. In contrast to bounds in the existing literature, the new bounds are straightforward and require little information about the signal to be quantized. The derivations of the new bounds provide motivation for new analysis of dithered quantizer performance that occurs later in this chapter. An analysis of dithering with a periodic noise source is presented in Section 2.6. The periodic noise source is considered because of its similarity to implementations involving linear feedback shift registers (LFSRs), or Pseudo-Noise (PN) generators. New analysis of phase dithering effects is presented in Sections 2.7 and 2.8, followed by simulation results and design examples.

2.2 Quantizer Model

When a discrete-time input signal, $x[n]$, passes through an ideal uniform mid-tread quantizer [8], the output signal, $y[n]$, can always be expressed as $y[n] = x[n] + e[n]$ where $e[n]$ is the quantization error. The quantization error is a deterministic function of $x[n]$. The input to the quantizer is mapped to one of 2^b levels, where b is the number of bits which digitally represent the input sample. Output levels are separated by one quantizer step size, $\Delta = 2^{-b}$.

Throughout this chapter Δ_A will be used as the step size for amplitude quantization results, Δ_P will be used for phase quantization results, and Δ will be used if the result applies to both amplitude and phase quantization. Similar subscripting will be used on the quantization error.

The input/output relation of a mid-tread quantizer appears in Fig. 2. If the input does not saturate the quantizer then the quantizer error is [8]:

$$e[n] = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} (-1)^k \frac{\Delta}{j2\pi k} \exp(\frac{j2\pi k x[n]}{\Delta}) \quad . \tag{2.1}$$

Note that the above equation is not correct when the quantizer input modulo Δ is $\frac{\Delta}{2}$. This is not a problem since future analysis involving this expression will treat the input as having a piecewise continuous probability density function in which case the collection of points in question have probability measure zero. If the input signal is bounded so that $|x[n]| \leq A_Q$ where $A_Q = 1/2 - \Delta$, then the quantizer does not saturate and $|e[n]| \leq \Delta/2$. Throughout this chapter, quantizers are always operating in non-saturation mode.

2.3 Amplitude Quantization Effects

Let a discrete-time sinusoid with amplitude $A \leq A_Q$ and frequency ω_0 be the input to a mid-tread quantizer. If the sinusoid is generated in a synchronous discrete-time system, ω_0 can be expressed as 2π times the ratio of two integers. The input sequence is then periodic. Since the error sequence, $e_A[n]$, is a deterministic function of the input sequence, it is periodic as well. Therefore, the spectrum of the error sequence will consist of discrete frequency components (spurs) which contaminate the spectrum of $x[n]$.

The following argument leads to an upper bound on the size of the largest frequency component in the spectrum of $e_A[n]$. Assuming the quantizer is not saturated by the input signal $x[n]$, the maximum possible quantization error is $\Delta_A/2$, where Δ_A is the amplitude quantization step size. The total power in $e_A[n]$ is then bounded by $\Delta_A^2/4$. By Parseval's relation, the sum of the spur powers in the spectrum of $e_A[n]$ equals the power in $e_A[n]$. In

order to maximize the power in a given spur, the total number of spurs must be minimized. Since $e_A[n]$ is real, the maximum power in a spur occurs when there are two frequency components at $+\omega_{spur}$ and $-\omega_{spur}$, with equal power[1]. With two frequency components the power in a single spur is $\leq \Delta_A^2/8$.

Since $x[n]$ is real, its spectrum consists of a positive and a negative frequency component, each having power $A^2/4$. Using the above bound on spur power, the Spurious-to-Signal Ratio ($SpSR$) is $\leq \Delta_A^2/(2A^2)$. If $A = A_Q \approx 1/2$ provided b is not small, then in deciBels with respect to the carrier (dBc), $SpSR \leq 3 - 6b$ dBc , where $\Delta_A = 2^{-b}$, and b is the word length in bits. In summary, this upper bound on power in a spur caused by amplitude quantization exhibits -6 dBc per bit behavior.

2.4 Phase Quantization Effects

Let a phase waveform, $\phi[n]$, be the input to the mid-tread quantizer. The phase waveform, $\phi[n] = \langle fn + \Phi/2\pi \rangle$ is a sampled sawtooth with amplitude ranging from 0 to 1. The frequency to be generated, measured in cycles/sample, is f, and the static phase, measured in radians, is Φ. Throughout the remainder of this chapter, "phase" refers to the instantaneous phase waveform, $\phi[n]$, unless explicitly stated otherwise. The fractional operator, $\langle x \rangle$, is defined so that $\langle x \rangle = x \mod 1$, e.g., $\langle 1.3 \rangle = 0.3$. Since $\phi[n]$ is generated by a synchronous, finite-word-length, discrete-time system, it has a finite period. The signal output from the quantizer can be expressed as $\phi[n] + e_P[n]$, where $e_P[n]$ represents the error introduced by quantization. Since $\phi[n]$ is periodic, $e_P[n]$ is periodic with a period less than or equal to the period of $\phi[n]$. After multiplication by 2π and passage through the ideal function generator, the output signal is $y[n] = A \cos(2\pi\phi[n] + 2\pi e_P[n])$. If the quantizer has many levels, i.e., > 16, $e_P[n] \ll 1$, and the small angle approximation $y[n] \approx A \cos(2\pi\phi[n]) - 2\pi A e_P[n] \sin(2\pi\phi[n])$ may be used.

Since $e_P[n]$ and $\phi[n]$ are periodic, the total error $2\pi A e_P[n] \sin(2\pi\phi[n])$ is periodic. The

[1]DC offsets and half sampling rate spurs are excluded because they can be corrected by appropriate calibration and filtering.

total error power is bounded by $\pi^2 A^2 \Delta_P^2$ because $e_P[n]$ is bounded by $\Delta_P/2$ and the magnitude of a sinusoid is bounded by unity. Recalling the arguments in the previous section on amplitude quantization effects, the maximum spur power of the real error signal is bounded by placing the total error power into two spectral components. Therefore, the maximum spur power is $\pi^2 A^2 \Delta_P^2/2$, where $\Delta_P = 2^{-b}$ and b bits are used to represent phase samples. By the above approximation for $y[n]$ and the bound on the spur power, the Spurious-to-Signal Ratio bound is $SpSR \leq 2\pi^2 \Delta_P^2 = 13 - 6b$ dBc, independent of the signal amplitude, A. This new, simple derivation demonstrates the underlying -6dBc per phase bit behavior, without the analytical complexity found in other existing bounds. More complicated arguments [7] improve the bound by about 9 dB.

2.5 Amplitude Dithering

In this section, rounding the sum of an already quantized sinusoid and an appropriate dither signal is shown to cause spurious magnitudes which depend on the original (longer) word length, not the output (shorter) word length. This phenomenon occurs at the expense of increased system noise from the addition of the dithering signal. An important finite word-length dithering system is subsequently shown to be equivalent to the continuous-amplitude uniformly-dithered system.

Consider the conceptual block diagram for a waveform generator shown in Fig. 3. The b-bit quantizer can be split into two parts as in Fig. 4: a high-resolution B-bit quantizer ($B > b$) followed by truncation or rounding to b bits. Thus, the generation process consists of two separate steps: production of a high-resolution waveform and reduction of the word length. The number of bits used to represent the high-resolution samples should be sufficient to guarantee the desired spectral purity. Then, the word length should be reduced without creating excess signal-dependent quantization error.

The input in Fig. 5 is a B-bit representation of a sinusoid, $x[n] = A\sin(2\pi\phi[n]) + e_{A0}[n]$, where $e_{A0}[n]$ is the quantization error. The dither signal, $z_u[n]$, is white noise uniformly

distributed over the interval $[-\Delta_A/2,\ \Delta_A/2)$, where $\Delta_A = 2^{-b}$. The sum $z_u[n] + x[n]$ is rounded to retain only the b most significant bits. The rounding can be modeled as a uniform quantizer with step size Δ_A. The amplitude A is chosen to avoid saturating this quantizer when the dither signal is added, i.e., $A + \Delta_A/2 \leq A_Q$.

The output from the quantizer can be expressed as $y[n] = x[n] + z_u[n] + e_A[n]$. The characteristic function of the dither signal, $z_u[n]$, is:

$$F_z(\alpha) = E\left\{\exp(j\alpha z[n])\right\} = \frac{2\sin(\alpha\Delta_A/2)}{\alpha\Delta_A} = \text{sinc}\left(\frac{\alpha\Delta_A}{2\pi}\right), \tag{2.2}$$

which has zeros at non-zero integer multiples of $2\pi/\Delta_A$. Thus, as shown in [1], $e_A[n]$ will be a white, wide-sense stationary process, uniformly distributed over $[-\Delta_A/2, \Delta_A/2)$. Further, $e_A[n]$ will not contribute spurious harmonics to the output spectrum of $y[n]$. Any spurious components in $y[n]$ are therefore due to $e_{A0}[n]$, which are present in the B-bit input to the quantizer, $x[n]$.

It remains to comment on the noise power not isolated in discrete spurious frequency components. Gray and Stockham have shown [13] that the power in the signal $y[n] - x[n] = e_A[n] + z_u[n]$ is $\Delta^2/6$. This is approximately twice the error variance of a quantization system with no dithering signal. The result is approximate because, in the absence of dither ($z_u[n] = 0$), the variance of $e_A[n]$ is generally not exactly $\frac{\Delta_A^2}{12}$. In summary, $y[n]$, which is quantized to b bits, exhibits spurious performance as if it was quantized to B bits ($B > b$), at the expense of approximately doubling the noise power.

Because the input $x[n]$ is expressed as a B-bit value, an important system equivalent to a system with continuous-amplitude, uniformly-dithered word-length reduction can be constructed. Replace the uniformly distributed dither signal, $z_u[n]$, by a finite word-length representation of it, $z[n]$, which is said to be discretely and evenly distributed over the $(B - b)$-bit quantized values in the interval $[-\Delta_A/2, \Delta_A/2)$. Heuristically, $z[n]$ randomizes the portion of the finite word-length input, $x[n]$, that is about to be thrown away by the rounded truncation. This process is equivalent to continuous uniform dithering, since if $x[n]$ is padded out to an infinite number of bits by placing zeros beyond the least significant bit (LSb), then only the $B - b$ most significant bits of $z_u[n]$ will have an effect on the resulting

sum, $x[n]+z_u[n]$. All of the bits below the most significant $B-b$ are added to zero, and cannot beget a carry. The output, $y[n]$, is identical in both systems. Therefore $z_u[n]$, continuously, uniformly distributed over $[-\Delta_A/2, \Delta_A/2)$ can be replaced by the discretely valued $z[n]$, and yield the same spurious response for $y[n]$.

It appears that the finite word-length dither signal, $z[n]$, could be generated by a linear feedback shift register (LFSR), or PN generator. This will be strictly true only if the PN generator has an infinite period, since, at this time, the dither signal is required to be white. However, it is not surprising that ideal behavior is approached as the period of the PN generator gets longer. With a sufficiently long period, the case where spur magnitudes are limited by the original word length can be achieved. The following section gives a simple model for a system implementation using a periodic random sequence which can be approximated by a PN generator.

2.6 The Effect of Periodic Dither

This section analyzes the use of a periodic dither signal with a long period, L, for both amplitude and phase dithering. Since the dither signal is periodic, the discrete frequency components in its spectrum will contaminate the desired signal. It is shown that the period can be chosen to satisfy worst case spurious specifications. In this section, the case where the dither signal is generated using one uniform variate ($M = 1$) is given. When the dither signal is the sum of M independent uniform variates ($M > 1$), as in Section 2.7, the analysis is the same because the resulting signal is an i.i.d. sequence of random variables.

Instead of using the white dither process, $z_u[n]$, described in the previous section, consider a substitute, $z_L[n]$, which is periodic with period L. Any two samples, $z_L[n]$ and $z_L[n+m]$, where $m \neq 0 \bmod L$, are independent. Samples of $z_L[n]$ are uniformly distributed between $[-\Delta/2, \Delta/2)$, and the quantization step size is Δ.

When $z_L[n]$ is used as the dither signal, let the quantizer error be called $e_L[n]$. The autocorrelation of $z_L[n]$ when the lag, m, is an integer multiple of L is equal to $R_{z_L z_L}[0] = \Delta^2/12$.

In the PN generator approximation to this noise source, $L = 2^l - 1$ where l is the length of the shift register in bits. At other lag values, the samples of $z_L[n]$ are independent, and since they have zero mean, the autocorrelation is zero. Therefore, expressing the autocorrelation as a discrete-time Fourier series:

$$R_{z_L z_L}[m] = \frac{\Delta^2}{12}\delta[m \bmod L] = \sum_{l=0}^{L-1} \frac{\Delta^2}{12L} \exp(\frac{j2\pi ml}{L}) \tag{2.3}$$

where $\delta[m]$ is the Kronecker delta function ($\delta[0] = 1$, $\delta[m] = 0$ for $m \neq 0$), and $z_L[n]$ contains L discrete frequency components, each with power $\Delta^2/(12L)$.

In the autocorrelation expression for $e_L[n]$, the expectation is taken over the random variables $z_L[n]$ and $z_L[n+m]$, using the definition of the autocorrelation and Equation 2.1:

$$R_{e_L e_L}[n, n+m] = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \alpha_k[n]\alpha_l^*[n+m] E\left\{\exp(\frac{j2\pi}{\Delta}(kz_L[n] - lz_L[n+m]))\right\} \tag{2.4}$$

where:

$$\alpha_k[n] = \frac{\Delta(-1)^k}{j2\pi k} \exp(\frac{j2\pi k s[n]}{\Delta}) \quad . \tag{2.5}$$

The desired signal to which the dither signal $z_L[n]$ is added is $s[n]$. Using the notation from earlier sections, in phase quantization, $s[n] = \phi[n]$ and in amplitude quantization $s[n] = x[n]$. When the lag is not a non-zero integer multiple of L,

$$\begin{aligned} E\left\{\exp(\frac{j2\pi}{\Delta}(kz_L[n] - lz_L[n+m]))\right\} &= E\left\{\exp(\frac{j2\pi kz_L[n]}{\Delta})\right\} E\left\{\exp(\frac{-j2\pi lz_L[n+m]}{\Delta})\right\} \\ &= F_z(\frac{2\pi k}{\Delta})F_z(\frac{-2\pi l}{\Delta}) = \delta[k]\delta[l]\,. \end{aligned}$$

This last fact is true because the characteristic function of $z_L[n]$ has zeros at all non-zero integer multiples of $2\pi/\Delta$ (Equation 2.2).But since k and l never assume the value 0 in Equation 2.4,the autocorrelation function is zero when the lag is not 0 mod L. When the lag is 0 mod L:

$$E\left\{\exp(\frac{j2\pi}{\Delta}(kz_L[n] - lz_L[n+m]))\right\} = E\left\{\exp(\frac{j2\pi(k-l)z_L[n]}{\Delta})\right\} = \delta[k-l] \quad .$$

This results in:

$$R_{e_L e_L}[n, n+m] = \frac{\Delta^2}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \cos\left(\frac{2\pi k}{\Delta}(s[n] - s[n+m])\right) \quad . \tag{2.6}$$

Setting $m = 0$ above and evaluating the resulting summation [9, page 7] yields the power in $e_L[n]$: $R_{e_L e_L}[n, n] = \Delta^2/12$. From Equation 2.6, $e_L[n]$ is a cyclo-stationary process because $s[n]$ has a finite period, N. Using the results of Ljung [10], spectral information is obtained when Equation 2.6 is averaged over time. Note that when the lag, m, is not only an integer multiple of L, the period of the dither, but also an integer multiple of N, the autocorrelation function equals $\Delta^2/12$, independent of n. The smallest non-zero lag that satisfies these two conditions is the least common multiple of L and N, denoted by qL where q is an integer. Therefore, the period of the time-averaged autocorrelation function, $\bar{R}_{e_L}[m] = \text{Avg}_n(R_{e_L e_L}[n, n+m])$, is at least L and at most qL. Let the period equal cL, where c is an integer, $1 \leq c \leq q$. The function $\bar{R}_{e_L}[m]$ can be expressed as a sum of cL weighted complex exponentials:

$$\bar{R}_{e_L}[m] = \sum_{l=0}^{cL-1} p_l \exp(\frac{j2\pi ml}{cL}), \quad m = \ldots, -1, 0, 1, 2, \ldots$$

where

$$p_l = \frac{1}{cL} \sum_{m=0}^{cL-1} \bar{R}_{e_L}[m] \exp(\frac{j2\pi ml}{cL}) = \frac{1}{cL} \sum_{n=0}^{c} \bar{R}_{e_L}[nL] \exp(\frac{j2\pi nl}{c}) \quad .$$

The last equality is true since the autocorrelation function in Equation 2.4 and its time-average, $\bar{R}_{e_L}[m]$ are zero for lags not equal to integer multiples of L. The weights, p_l,$l = 0, 1, \ldots cL-1$, are the power magnitudes of the spurs. Since $\bar{R}_{e_L}[m] \leq \Delta^2/12$, the spur power can be bounded: $p_l \leq \frac{\Delta^2}{12cL} \leq \frac{\Delta^2}{12L}$. Equality is achieved when the period of the time-averaged autocorrelation function is exactly L, the period of the dither.

As $L \to \infty$, the spacing between spurs goes to zero in the spectra of both $e_L[n]$ and $z_L[n]$. The power in an individual spur goes to zero, but the density (power per unit of frequency) tends to a constant. Thus, ideal white noise behavior is approached. While $z_L[n]$ and $e_L[n]$ are correlated in general, the worst case spur power scenario coherently adds the power spectra from both processes. For this reason, L should be chosen to satisfy $\Delta^2/(6L) < P_{\max}$, where $P_{\max}$ is the maximum acceptable spur power. When constructing a dither signal as the sum of $M \geq 1$ independent, uniform variates the noise autocorrelation becomes $R_{z_L z_L}[m] = (M\Delta^2/12)\delta[m \bmod L]$. The analysis follows closely to that for $M = 1$, and L should be chosen to satisfy $(M+1)\Delta^2/(12L) < P_{\max}$.

As in the previous section, since the desired signal has finite word length, it is equivalent to round or truncate the dither signal to an appropriate word length. An implementation using a PN generator is an approximation to such a truncated periodic noise source which produces a periodic sequence of discretely and evenly distributed random numbers.

2.7 Phase Dithering

In this section, phase dithering is analyzed using a continuous, wide-sense stationary sequence. As described in Section 2.5 on amplitude dithering, an evenly distributed discrete random sequence is equivalent to continuous uniform dithering when the initial phase word is quantized to a finite number of bits. The dither signal is constructed by summing M independent sequences of i.i.d. variates. Each variate is uniformly distributed over one quantization interval, $[0, \Delta_P)$, where $\Delta_P = 2^{-b_P}$, and b_P is the number of bits used to represent the phase after word-length reduction. The dither signal is referred to as an M^{th}-order dither signal because its characteristic function has M^{th}-order zeros at non-zero integer multiples of $\frac{2\pi}{\Delta_P}$.

It is desired to find the variance of the noise introduced by M^{th}-order phase dithering. Let $x[n] = \cos(2\pi\hat{\phi}[n])$ represent a digital sinusoid that has been generated by a system using M^{th}-order phase dithering. Amplitude quantization effects are ignored by assuming a sufficiently large number of bits are used to represent each amplitude sample. The normalized phase signal is written as $\hat{\phi}[n] = \phi[n] + \epsilon[n]$, where $\phi[n]$ is the ideal phase signal and $\epsilon[n]$ is the *quantization noise signal.* The ideal phase signal is defined in Section 2.4. The quantization noise signal is the sum of the dither signal, $z[n]$ and the quantization error signal, $e[n]$, defined in Equation 2.1. The normalized phase sequence is the digital signal from which the system constructs the amplitude sample sequence, $x[n]$. The normalized phase signal is generated by quantizing the sum of the ideal phase signal and an M^{th}-order dither signal, $z[n]$. The dither signal, $z[n]$, is an i.i.d sequence constructed by summing M independent variates, each uniformly distributed over one quantization interval, $[0, \Delta_P)$. The quantization interval is defined by the quantizer that the normalized phase sequence, $\hat{\phi}[n]$, emerges from. This

type of a quantizer system is called a "non-subtractive dither system" and was studied in detail by Gray and Stockham [13].

Define the error due to phase dithering to be $e_x[n] = x[n] - \cos(2\pi\phi[n])$. Using small angle approximations:

$$e_x[n] = 2\pi\epsilon[n]\sin(2\pi\phi[n]) + O(\epsilon^2[n]).$$

Since the dither signal and the quantization error are $O(\Delta_P)$ it follows that their sum, $\epsilon[n]$, is $O(\Delta_P)$. The phase quantization interval, Δ_P, is assumed to be small. The autocorrelation function of the error due to phase dithering is $R_{e_x e_x}[n,m] = E\{e_x[n]e_x[m]\}$, where $E\{\}$ denotes the expectation operator:

$$R_{e_x e_x}[n,m] = 4\pi^2 \sin(2\pi\phi[n])\sin(2\pi\phi[m])E\{\epsilon[n]\epsilon[m]\} + O(\Delta_P^3).$$

The following variance discussion assumes that the dither signal is the sum of two or more ($M \geq 2$) uniform variates. The variance of the first-order dithering system ($M = 1$) has been bounded using different techniques [11], and the final bound will be cited after this analysis. Note that the following analysis does not bound the variance, but rather gives an expression for it that is valid for small quantization step sizes.

For $M > 1$, Gray and Stockham [13] showed that the quantization noise signal, $\epsilon[n]$, is white with variance $\frac{(M+1)\Delta^2}{12}$. Therefore, we set $E\{\epsilon[n]\epsilon[m]\} = \frac{(M+1)\Delta^2}{12}\delta[n-m]$ where $\delta[m]$ is the Kronecker delta function ($\delta[0] = 1$, $\delta[m] = 0$, $m \neq 0$). We average the above autocorrelation function over time [10] in order to obtain spectral information. Assuming the desired output frequency is neither 0 nor π, the time averaged value of $\sin^2(2\pi\phi[n])$ is 0.5. This is an excellent assumption because systems using digital sinusoid generators typically do not require these two pathological frequencies [15]. As the phase quantization interval, Δ_P, becomes small, the order term can be dropped and the variance of the error due to phase dithering is $\frac{(M+1)\pi^2\Delta^2}{6}$. In the absence of order terms, the noise is white and therefore often tolerable in wide-band systems.

Since we assume that we do not generate the frequencies 0 or π, the desired signal power is 0.5, and the signal-to-noise-ratio (SNR) is $\frac{3}{\pi^2(M+1)\Delta_P^2}$ or $6.02b_P - 10\log_{10}(M+1) - 5.17$ dBc

when $\Delta_P = 2^{-b_P}$ and b_P phase bits are left after quantizing the phase. The noise power spectral density (NPSD) is another important frequency synthesizer metric. As its name implies, the NPSD measures the amount of noise per unit Hz as a function of frequency. In this discrete-time, phase-dithered system, the noise power is roughly uniform across the frequency interval $(-\pi, \pi)$. This frequency span corresponds to f_s, the digital clock frequency. We therefore scale the noise variance by the digital clock frequency, f_s, and the signal power in order to obtain the NPSD relative to the carrier for $M > 1$:

$$NPSD = 5.17 + 10\log_{10}(M+1) - 6.02b_P - 10\log_{10}(f_s) \text{ dBc/Hz}, \tag{2.7}$$

independent of frequency. For the special $M = 1$ case, [11] bounded the noise power spectral density for the special $M = 1$ case as $NPSD_{M=1} \leq 9.94 - 6.02b_P - 10\log_{10}(f_s)$ dBc/Hz. Some values of the NPSD are given in Table III.

The spurious performance of the phase dithered system is found by considering the autocorrelation function of the digital sinusoid, $E\{x[n]x[n+m]\}$ when the lag, m, is non-zero. When this is the case, $x[n]$ and $x[n+m]$ are independent because the dither signal, $z[n]$, is i.i.d. The autocorrelation function becomes the product of two expectations and it suffices to consider the "expected waveform," $E\{x[n]\}$, to obtain spectral information. Since the dither signal and the quantization error are $O(\Delta_P)$ and Δ_P is small in practice, we can expand the output, $x[n]$, in a Taylor series expansion about $\epsilon[n] = 0$:

$$x[n] = \sum_{k=0}^{\infty} \frac{(2\pi)^k \epsilon^k[n] f^{(k)}(2\pi\phi[n])}{k!},$$

where $f^{(k)}(\theta)$ denotes the k^{th} derivative of the function $f(\theta) = \cos(\theta)$. The k^{th} term in the expansion is proportional to the quantization noise signal raised to the k^{th} power. To evaluate the expected waveform, we take the expected value of the Taylor series:

$$E\{x[n]\} = \cos(2\pi\phi[n]) + \sum_{k=1}^{\infty} \frac{(2\pi)^k E\{\epsilon^k[n]\} f^{(k)}(2\pi\phi[n])}{k!}. \tag{2.8}$$

It was shown by Gray and Stockham [13] that the first M moments of the quantization noise, $\epsilon[n]$, are independent of the ideal phase signal, $\phi[n]$, when the phase dither signal is

the sum of M independent uniform variates. Therefore, each of the first M terms in the above sum over k can be written as a constant times either $\cos(2\pi\phi[n])$ or $\sin(2\pi\phi[n])$. Since each of these terms has frequency support only where the desired sinusoid, $\cos(2\pi\phi[n])$, has support, these terms do not generate spurious harmonics. The first M terms make the effective amplitude of the ideal portion of the output signal, $x[n]$, different from unity. The difference is small, however, because the largest of the first M terms is $O(\Delta_P)$.

The $(M+1)^{th}$ moment of the quantization noise will generally be dependent on the ideal phase signal, $\phi[n]$, which is a periodic sawtooth signal defined in Section 2.4. Therefore, the $(M+1)^{th}$ moment is generally periodic and gives rise to spurs which are scaled by $\frac{(2\pi)^{M+1}}{(M+1)!}$ and amplitude modulated by the sinusoid $f^{(M+1)}(2\pi\phi[n])$. While higher-order moments will also contribute spurs, these will be small compared to the $(M+1)^{th}$ term in the expected waveform when the quantization interval, Δ_P, is small. The nature of the periodicity of the $(M+1)^{th}$ moment is studied in the next section. Since $\epsilon[n]$ is $O(\Delta_P)$, its $(M+1)^{th}$ moment is $O(\Delta_P^{(M+1)})$. Therefore, the maximum spur power is $O(\Delta_P^{2(M+1)})$, giving the $6(M+1)$ dBc per phase bit result. In order to tighten the bound on the maximum spur power, the following section considers the constant term in the order expression $O(\Delta_P^{2(M+1)})$.

2.8 Detailed Phase Spur Bounds

The previous section showed that the spurs in an M^{th}-order phase dithering system are due to the general dependence of the $(M+1)^{th}$ quantization noise moment, $E\{\epsilon^{M+1}[n]\}$, on the periodic, ideal phase signal, $\phi[n]$. This section investigates the behavior of the moment as a function of the order of the phase dithering, M, and the desired phase input, $\phi[n]$. Worst-case spur bounds are obtained and are shown to be achieved for large classes of output frequencies.

Zimmerman's contribution to [11] included an expression for the second moment of the quantization noise in a first-order phase dithering system. The second moment of the quantization noise was shown to be a polynomial in the quantity $p[n]$, where $p[n]$ is defined to

be the ideal phase modulo Δ_P, scaled by Δ_P, $p[n] = \langle \frac{\phi[n]}{\Delta_P} \rangle$. Zimmerman analyzed the polynomial and constructed a phase-spur bound that was achieved for a large class of output signals. For the special case of second-order phase dithering, [12] showed how the third moment of the quantization noise could also be expressed as a polynomial in $p[n]$. Another phase-spur bound was obtained but it was not shown to be achieved. This section extends Zimmerman's earlier work to dither orders up to $M = 5$ and emphasizes a graphical point of view. New phase-spur bounds are then derived for phase dithering orders $M = 2, 3, 4, 5$. These are shown to be practically achieved for large classes of output signals.

The quantization noise was defined in the last section to be $\epsilon[n] = \hat{\phi}[n] - \phi[n]$, where $\phi[n]$ is the ideal phase signal and $\hat{\phi}[n]$ is the phase signal emerging from the phase quantizer. In light of this last fact, it is instructive to write the quantization noise as $\epsilon[n] = Q(\phi[n] + z[n]) - \phi[n]$, where $z[n]$ is the dither signal and Q denotes the quantization operator. Without loss of generality, let us assume that the independent variates that are summed together to generate the dither signal, $z[n]$, are uniformly distributed over the interval $[0, \Delta_P)$, and that the quantizer is a truncator, i.e., $Q(x) = \Delta_P \lfloor \frac{x}{\Delta_P} \rfloor$. For finite M and a given time index, n, the dither sample $z[n]$ is a well-defined random variable. Since the dither process is the sum of M independent processes whose samples are i.i.d. variates, it follows that the p.d.f. of the sample $z[n]$, $f_M(z)$, is independent of the time index, n, and is the convolution of M uniform p.d.f.'s [14].

Given a sample of the ideal phase, $\phi[n]$, at some time instant, n, we may write all of the possible outcomes of $\epsilon[n]$ and their probabilities. Consider Fig. 6. When the dither signal is $0 \leq z[n] < \Delta_P - \phi[n] + Q(\phi[n])$, the output of the quantizer is $Q(\phi[n])$ and the quantization error is $Q(\phi[n]) - \phi[n]$. For $i = 1, 2, \cdots, M-1$, when the dither signal is $i\Delta_P + Q(\phi[n]) - \phi[n] \leq z[n] < (i+1)\Delta_P + Q(\phi[n]) - \phi[n]$, the output of the quantizer is $Q(\phi[n]) + i\Delta_P$ and the quantization error is $Q(\phi[n]) + i\Delta_P - \phi[n]$. Finally, when the dither signal is $M\Delta_P + Q(\phi[n]) - \phi[n] \leq z[n] < M\Delta_P$, the output of the quantizer is $Q(\phi[n]) + M\Delta_P$ and the quantization error is $M\Delta_P Q(\phi[n]) - \phi[n]$. It will be convenient to use the quantity $p[n] = \frac{1}{\Delta_P}(\phi[n] - Q(\phi[n]))$, which is the ideal phase modulo Δ_P, scaled by Δ_P, mentioned

earlier in this section. Note that this quantity ranges from zero to unity. We can write the outcomes of the quantization noise as:

$$\epsilon[n] = \begin{cases} -\Delta_P\, p[n] & \text{with probability } P_0, \\ \Delta_P - \Delta_P\, p[n] & \text{with probability } P_1, \\ \vdots & \vdots \\ i\,\Delta_P - \Delta_P\, p[n] & \text{with probability } P_i, \\ \vdots & \vdots \\ M\,\Delta_P - \Delta_P\, p[n] & \text{with probability } P_M, \end{cases}$$

where the probabilities are:

$$P_0 = \int_0^{\Delta_P(1-p[n])} f_M(z)dz,$$

$$P_i = \int_{\Delta_P(i-p[n])}^{\Delta_P(i+1-p[n])} f_M(z)dz \quad i = 1, 2, \cdots, M-1,$$

$$P_M = \int_{\Delta_P(M-p[n])}^{M\Delta_P} f_M(z)dz.$$

Therefore, the L^{th} quantization noise moment can be written as:

$$E\{\epsilon^L[n]\} = \Delta_P^L \sum_{i=0}^{M} (i - p[n])^L P_i.$$

It is easily shown by induction on M that $f_M(z)$ is a polynomial in z over the interval $\Delta_P k \leq z < \Delta_P(k+1)$ for integer k. Therefore, the above probabilities, P_i, can be written as polynomials in the variable $p[n]$ defined above. Finally, the L^{th} moment of the quantization noise can also be expressed as a polynomial in $p[n]$. Based on the analysis of the previous section, the phase-spur performance of the M^{th}-order phase dithered system can be evaluated by considering the case where $L = M+1$. The cases of $M = 1, 2$ have been reported already [11, 12]. After considerable algebra, the moments for $M = 3, 4, 5$ can be obtained and are presented in the following table.

Table I: Quantization Noise Moments Versus Phase Dither Order, M

M	$(M+1)^{th}$ Quantization Noise Moment
1	$\Delta_P^2(p[n] - p^2[n])$
2	$\frac{\Delta_P^3}{2}(1 - p[n] + 3p^2[n] - 2p^3[n])$
3	$\frac{\Delta_P^4}{3}(10 - 3p^2[n] + 6p^3[n] - 3p^4[n])$
4	$\frac{\Delta_P^5}{12}(303 + 2p[n] - 20p^3[n] + 30p^4[n] - 12p^5[n])$
5	$\frac{\Delta_P^6}{2}(455 + p^2[n] - 5p^4[n] + 6p^5[n] - 2p^6[n])$

These polynomials are plotted in Figs. 7 – 11. For any dither order, M, a worst-case phase spur bound can be obtained by determining the peak-to-peak variation of the $(M+1)^{th}$ moment of the quantization noise when $p[n] \in [0,1)$. This is because the moment in question could be written as a time-invariant offset plus $\pm\frac{\Delta_P^{M+1} S_{M+1}}{2}\cos(\pi n)$ where S_{M+1} is the peak-to-peak variation, or span, of the $(M+1)^{th}$ moment scaled by Δ_P^{M+1}. Then the largest spur-causing term in Equation 2.8 becomes:

$$\pm\frac{S_{M+1}(2\pi\Delta_P)^{M+1}}{2(M+1)!}\{\begin{array}{c}\sin\\ \cos\end{array}\}(2\pi\phi[n])\cos(\pi n).$$

This can also be written as

$$\pm\frac{S_{M+1}(2\pi\Delta_P)^{M+1}}{2(M+1)!}\{\begin{array}{c}\sin\\ \cos\end{array}\}(2\pi\phi[n]+\pi n).$$

Assuming that the power in the desired signal is $\frac{1}{2}$, it follows that the spur-to-signal ratio (SpSR) can be bounded by:

$$SpSR \leq (\frac{S_{M+1}}{2})^2\left(\frac{(2\pi\Delta_P)^{M+1}}{(M+1)!}\right)^2.$$

The peak-to-peak variations can be obtained from the analytic expressions for the moments in the above table. The SpSR is plotted in Fig. 12 and is presented in the following table in dBc:

Table II: Higher-Order Dithering Phase Spur Bounds

M	Phase Spur Bound (dBc)
1	$7.84 - 12.04b_P$
2	$5.97 - 18.06b_P$
3	$6.15 - 24.08b_P$
4	$6.00 - 30.10b_P$
5	$6.04 - 36.12b_P$

For the the case of first-order phase dithering, $M = 1$, it was noted in [11] that the worst-case spur bound was achieved by a large class of output signals. This is best illustrated graphically. In Fig. 7, the span of the second moment is 1/4. At time instant $n = 0$, let the initial, static, phase be $K_1\Delta_P + K_2\frac{\Delta_P}{2}$ where K_1 is any integer and K_2 is either 0 or 1.

Let the phase increment be $(K_3 + \frac{1}{2})\Delta_P$ for arbitrary integer K_3. Then $p[n] = \langle \frac{\phi[n]}{\Delta_P} \rangle$ will toggle back and forth between the values 0 and $\frac{1}{2}$. As a result, the second moment of the quantization noise will toggle back and forth between the values 0 and $\frac{1}{4}$, as seen in Fig. 7. Based on the discussion above for the worst-case spur bound, this achieves the bound since the second moment of the quantization noise is precisely a time-invariant constant plus an alternating term with period equal to two. Because of the generality of the constants K_1, K_2 and K_3, this worst-case scenario is achieved by a large class of output signals.

Consideration of the fourth and sixth quantization noise moments, shown in Figs. 9 and 11, respectively, shows that the preceding argument for $M = 1$ applies to $M = 3$ and $M = 5$ since these moments have their extrema at $p[n] = 0$ and $p[n] = 0.5$. Therefore, third-order and fifth-order phase dithering systems exactly achieve their worst-case spur bounds for a large class of output signals.

The third and fifth quantization noise moments exhibit different behavior than the preceding even moments. As shown in Figs. 8 and 10, the worst-case initial phase values modulo Δ_P can no longer be 0 or $\frac{\Delta_P}{2}$ since the third and fifth moments evaluated at $p[n] = 0$ equal the respective moment evaluated at $p[n] = 1/2$. For these moments, it is still desired to have the phase increment to be as it was above since then the third or fifth moment would be periodic with period two, which would place all of the error power in a single spur. The difference between the third moment evaluated at $p[n] + \frac{1}{2}$ and the third moment evaluated at $p[n]$ is maximized when $p[n] = \frac{1}{4}$. When the phase increment is $(K_3 + \frac{1}{2})\Delta_P$, where K_3 is any integer, $p[n]$ toggles back and forth between 1/4 and 3/4. The resulting peak-to-peak variation of the third moment is $\frac{3}{32} = 0.09375$ compared to the maximal peak-to-peak variation of approximately 0.09623. Therefore, an achievable spur-to-signal ratio for second-order phase dithering, $M = 2$, is only 0.23 dB different from the worst-case analysis above. Analysis of fourth-order phase dithering mirrors that of second-order phase dithering just given. When $p[n]$ is constrained to be on the interval $[0, 1)$, the difference between the fifth moment evaluated at $p[n]$ and the fifth moment evaluated at $p[n] + \frac{1}{2}$ is maximized when $p[0] = \frac{1}{4}$. With the same phase increment as above, the resulting peak-to-peak variation in the fifth

moment is 0.04887 compared to the maximal peak-to-peak variation of 0.04892. Therefore, an achievable spur-to-signal ratio for fourth-order phase dithering, $M = 4$, is less than 0.009 dB different from the worst-case analysis above.

2.9 Performance Comparison

The following comparison demonstrates the savings of higher-order phase dithering over no phase dithering and first-order phase dithering. It is desired to have better than -90 dBc spurious performance when the clock rate is 100 MHz. The results are detailed in Table III. The system with no phase dithering [7] requires 17 look-up phase bits, the system with first-order phase dithering requires 9 look-up phase bits, and the second-order phase dithering system requires only 6 look-up phase bits. The second-order phase dithering system memory consists of only 16 entries when quadrant symmetries are used. In contrast, the system with no phase dithering requires 32,768 entries. As expected, the noise power increases with the order of the dithering.

Table III: System Comparison for -90 dBc spurious and 100 MHz clock

Phase Dither Order	Look-up Phase Bits	Memory Entries	Noise Power Spectral Density
none	17	32768	-173.8 dBc/Hz
1	9	128	-124.2 dBc/Hz
2	6	16	-106.2 dBc/Hz

The noise power spectral density of the second-order phase-dithered system is higher than the phase noise density of economically priced analog frequency synthesizers. In addition, when used as part of a direct digital frequency synthesizer (DDS), the noise power produced by second-order dithering to 6 bits is greater than the quantization noise introduced by 8 bit digital-to-analog conversion, regardless of sampling rate. Therefore, second-order phase dithering down to 8 bits might be used in a high performance DDS system, resulting in phase-spur performance that vastly exceeds our example requirement, yet giving a table size

half that of the first-order dithered system that just meets the requirement. As a result, second-order phase dithering down to the minimum number of bits required is considered useful for high-rate applications in which coarse-resolution digital-to-analog converters, e.g., less than 5 bits, are used, or to reduce complexity in those applications where noise power spectral density requirements on the NCO are loose.

In applications where one can tolerate the level of noise introduced by using second-order phase dithering, it makes sense to amplitude dither the resulting sinusoid values to reduce the word length to be equal to the reduced phase word length. The amplitude dithering technique is the straightforward addition of a uniformly distributed $[0, \Delta)$ variate, as described in Section 2.5. This will add a white noise component of variance $\frac{\Delta_A^2}{12}$, which reduces the signal-to-noise ratio by only 0.27 dB. For the -90 dBc example given, a reduction from 16 amplitude bits to 6 amplitude bits would reduce the complexity of data path and input/output hardware by 62.5%. If the sinusoid is multiplied digitally, savings in multiplier complexity would be much greater. A digital phase-locked loop (DPLL) is one application where the phase noise of higher-order phase-dithered systems could be tolerated. This is because the narrow loop bandwidth of a DPLL can reject much of the phase noise of the sinusoid generator prior to output.

2.10 Simulation Results

Simulations were performed to validate the results of this analysis. These results were obtained using 8192-point unwindowed FFTs, and the synthesized frequencies were chosen to represent worst-case amplitude and phase spur performance. In each of the figures, ten power spectra were averaged to better show the spurious content of the signals. Fig. 13 shows the power spectrum of a sine wave of one-eighth the sampling frequency truncated to 8 bits of amplitude without dithering. Fig. 14 shows the same spectrum with a sixteen-bit sinusoid amplitude dithered with one uniform variate prior to truncation to 8 bits. Note that the spurs have been eliminated to the levels consistent with those imposed by the initial sixteen

bit quantization.

Fig. 15 shows the spectrum of a 5-bit phase-truncated sinusoid with high-precision amplitude values. A worst-case example of first-order phase dithering is shown in Fig. 16. The measured noise power spectral density in Fig. 16 is -62.3 dBc per FFT bin, giving a noise density of $-23.2 - 10\log_{10}(f_s/2)$ dBc, in agreement with the upper bound in [11]. The spur level is -52.3 dBc in the first-order dithered Fig. 16.

Fig. 17 shows the same example using second-order ($M = 2$) dithering using the sum of two uniform deviates. While the spectrum in Fig. 16 shows the residual spurs at -12 dBc per bit due to second-order effects, Fig. 17 shows no visible spurs, indicating better than -63 dBc spurious performance. Additional simulations involving Megapoint FFTs and not represented by figures confirm the -18 dBc per bit performance of the second-order phase-dithered system.

Fig. 18 shows a worst-case result for first-order phase dithering together with first-order amplitude dithering. The amplitude samples are truncated to 8 bits, as are the phase samples. Note that the spurs are not visible in the spectrum; however, close analysis has demonstrated that they are present at the -88 dBc level expected due to second-order effects.

Finally, Fig. 19 shows a worst-case result for second-order dithering to 4 bits of phase. The uniform variates that comprised the phase dither signal were generated using two 24-tap LFSR PN generators with the same polynomial. Multi-bit words were formed from the instantaneous contents of the shift registers. As such, the dither sequences are not quite white, with a slight low-pass coloration and less than 5 dB drop across the entire band. In order to show spurs clearly, a 16 Megapoint FFT was performed in this case. For display purposes, the positive frequency half the resulting spectrum was compressed by taking the maximum of each block of 512 adjacent power values. The results show residual second-order spurs at the predicted levels for 4 bits, approximately -72 dBc.

2.11 A First-Order Dithering Design Example

The block diagram of a direct digital frequency synthesizer using first-order phase dithering is shown in Fig. 20. The following system would perform at a sampling rate of 160 MHz, producing 8-bit digital sinusoids spur-free to -90 dBc with better than -120 dBc/Hz noise power spectral density. The system parameters are as follows:

Phase bits are in unsigned fractional cycle representation with:

phase accumulator word-length determined by frequency resolution, and

$\geq$ 16 bits prior to addition of 1 uniform phase dither variate, with $\geq$ 9 bits after dither addition and truncation;

Amplitude look-up-table with:

$\geq 2^7 = 128$ entries (using quadrant symmetries) of $\geq$ 16 bits each normalized so that the sinusoid amplitude equals 512 16-bit quantization steps less than the full-scale value;

Linear feedback shift register PN generator with $\geq$ 16 lags producing one 8-bit amplitude dither variate, and

One LFSR PN generator with $\geq$ 18 lags for generation of the 7-bit phase dither variate.

2.12 A Second-Order Dithering Design Example

The block diagram of a digital sinusoid generator incorporating second-order dithering is shown in Fig 21. The dramatic reduction in look-up table size and arithmetic complexity enabled by the use of second-order dithering makes this -90 dBc spur-free NCO easily implementable in VLSI. The second-order phase dither produces a digital sinusoid contaminated with periodic noise, white to a level of -90 dBc, at a signal-to-noise ratio of 20 dB. Such a digital sinusoid generator is useful as a local oscillator in high-rate (> 100 MHz) applications or in noisy signal environments.

The system parameters are as follows:

Phase bits are in unsigned fractional cycle representation with:

phase accumulator word-length determined by frequency resolution, and

$\geq$ 16 bits prior to addition of the sum of 2 uniform phase dither variates, with $\geq$ 5 bits after dither addition and truncation;

Amplitude look-up-table with:

$\geq$ 8 entries (using quadrant symmetries) of $\geq$ 16 bits each normalized so that the sinusoid amplitude equals 2048 16-bit quantization steps less than the full-scale value;

Linear feedback shift register PN generator with $\geq$ 18 lags producing one 11-bit amplitude dither variate, and

Two LFSR PN generators each with $\geq$ 19 lags for generation of two 12-bit phase dither variates, initialized to different states.

A 12-bit full adder to sum the two phase-dither variates into a 13-bit result.

2.13 Conclusion

A digital dithering approach to spur reduction in the generation of digital sinusoids has been presented. A class of periodic dithering signals has been analyzed because of its similarity to LFSR PN generators.

The advantage gained in amplitude dithering provides for spur performance determined by the longer pre-quantization word length when the digital dithering signal is an i.i.d. sequence distributed evenly, not uniformly, over one quantization interval. The reduced word length allows the use of less complicated multipliers and narrower data paths in purely digital applications. If the waveform is ultimately converted to an analog value, the reduced word length allows the use of fast, coarse-resolution, highly-linear digital-to-analog converters

(DACs) to obtain sinusoids or other periodic waveforms whose spectral purity is limited by the DAC linearity, not its resolution. These results suggest that coarsely quantized, highly-linear techniques for digital-to-analog conversion such as delta-sigma modulation would be useful in direct digital frequency synthesis of analog waveforms.

The advantage gained in the proposed method of phase dithering provides for an acceleration beyond the normal 6 dB per bit spur reduction to a $6(M+1)$ dB per bit spur reduction when the dithering signal consists of M uniform variates. Often the most convenient way to generate a periodic waveform is by table look-up with a phase index. Since the size of a look-up table is exponentially related to the number of phase bits, this can provide a dramatic reduction in the complexity of numerically-controlled oscillators, frequency synthesizers, and other periodic waveform generators.

The advantages of dithering come at the expense of an increased noise content in the resulting waveform. However, the noise energy is spread throughout the sampling bandwidth. While the noise level added in taking full advantage of even second-order dithering ($M = 2$) may be too high for low speed, high SNR applications, the whiteness of the added noise makes this technique ideal for use in high speed (> 500 MHz) integrated-circuit sinusoid generators. Linear feedback shift register-based PN generators can be used as a low-complexity dither generator with this technique, providing a straightforward path to VLSI implementation. For further circuitry savings, the resulting sinusoid samples may be amplitude dithered to reduce their word length without significant loss.

References

1. A.B. Sripad and D.L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Transactions on Acoustics Speech and Signal Processing* , vol. ASSP-25, pp. 442-448, October 1977.

2. L. Schuchman, "Dither signals and their effects on quantization noise," *IEEE Transactions on Communication Technology,* vol. COM-12, pp. 162-165, December 1964.

3. N.S. Jayant and L.R. Rabiner, "The application of dither to the quantization of speech signals," *Bell System Technical Journal,* vol. 51, pp. 1293-1304, July-August 1972.

4. S.C. Jasper, "Frequency Resolution in a Digital Oscillator," U.S. Patent Number 4,652,832, March 24, 1987.

5. P. O'Leary and F. Maloberti, "A Direct-Digital Synthesizer with Improved Spectral Performance," *IEEE Transactions on Communications,* vol. COM-39, pp. 1046-1048, July 1991.

6. J. Tierney, C. Rader, B. Gold, "A Digital Frequency Synthesizer," *IEEE Transactions on Audio and Electroacoustics,* vol. AU-19, No. 1, pp. 48-57, March 1971.

7. T. Nicholas and H. Samueli, "An Analysis of the Output Spectrum of Direct Digital Frequency Synthesizers in the Presence of Phase Accumulator Truncation," *41st Annual Frequency Control Symposium,* USERACOM, Ft. Monmouth, NJ, pp. 495-502, 1987.

8. L.E. Brennan and I.S. Reed, "Quantization Noise in Digital Moving Target Indication Systems," *IEEE Transactions on Aerospace and Electronic Systems,* vol. AES-2, pp. 655-658, November 1966.

9. I.S. Gradsteyn and I.M. Rizhk, Table of Integrals, Series, and Products, Academic Press, New York, corrected and enlarged edition, 1980.

10. L. Ljung, System Identification: Theory for the User, Prentice-Hall, Englewood Cliffs, NJ, 1987.

11. M.J. Flanagan and G.A. Zimmerman, "Spur-Reduced Digital Sinusoid Generation," to appear, *IEEE Transactions on Communications.*

12. M.J. Flanagan and G.A. Zimmerman, "Spur-Reduced Digital Sinusoid Generation Using Higher-Order Phase Dithering," Conference Record: Papers Presented at the 27th

Annual Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, IEEE Computer Society Press, November 1993.

13. R.M. Gray and T.G. Stockham, Jr., "Dithered Quantizers," *IEEE Transactions on Information Theory,* vol. IT-39, pp. 805-812, May 1993.

14. W. Feller, An Introduction to Probability Theory and Its Applications, Volume II, Wiley, New York, 1966.

15. R. J. Zavrel, Jr. and G. Edwards, The DDS Handbook: Second Edition, Stanford Telecommunications, Santa Clara, CA, 1990.

Figures

1. Spur generation in conventional digital sinusoid generation
2. Input/output relation of a midtread quantizer
3. Conceptual waveform generator model
4. Two-step waveform generator model
5. Uniform dithered quantizer
6. Phase quantization
7. First-order dithering: quantization noise second moment
8. Second-order dithering: quantization noise third moment
9. Third-order dithering: quantization noise fourth moment
10. Fourth-order dithering: quantization noise fifth moment
11. Fifth-order dithering: quantization noise sixth moment

12. Worst-case spur bounds vs. dither order, M, and phase bits

13. Power spectrum of 8 sample/cycle sine wave without dithering (8 bit amplitude quantization)

14. Power spectrum of 8 sample/cycle sine wave with amplitude dithering (8 bit amplitude quantization)

15. Power spectrum of 5-bit phase-truncated sine wave without phase dithering (high-precision amplitude)

16. Power spectrum of 5-bit phase-truncated sine wave with first-order phase dithering (high-precision amplitude)

17. Power spectrum of 5-bit phase-truncated sine wave with second-order phase dithering (high-precision amplitude)

18. Worst-case power spectrum of sinusoid with first-order phase dithering and amplitude dithering (8 bits each)

19. Worst-case power spectrum of sinusoid with second-order phase dithering to 4 phase bits (high-precision amplitude)

20. Block diagram of a first-order phase dithering system

21. Block diagram of a second-order phase dithering system

Fig. 1: Spur generation in conventional digital sinusoid generation

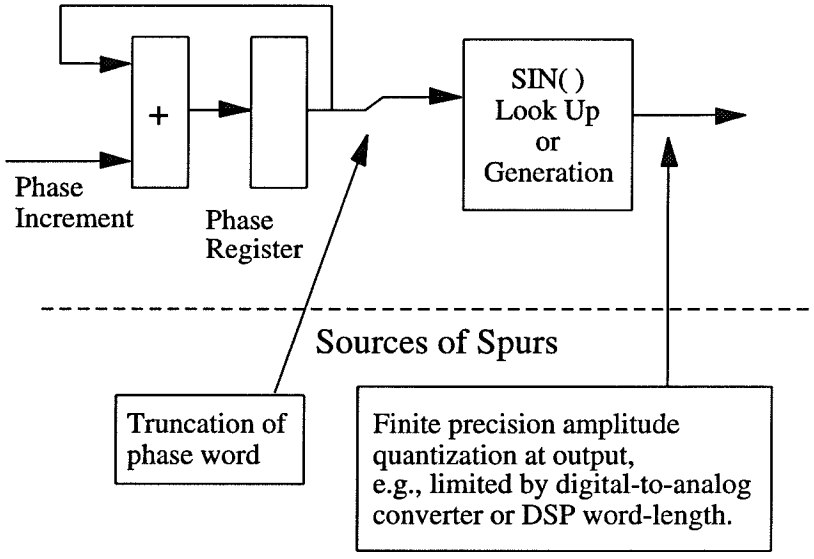


Fig. 2: Input/output relation of a midtread quantizer

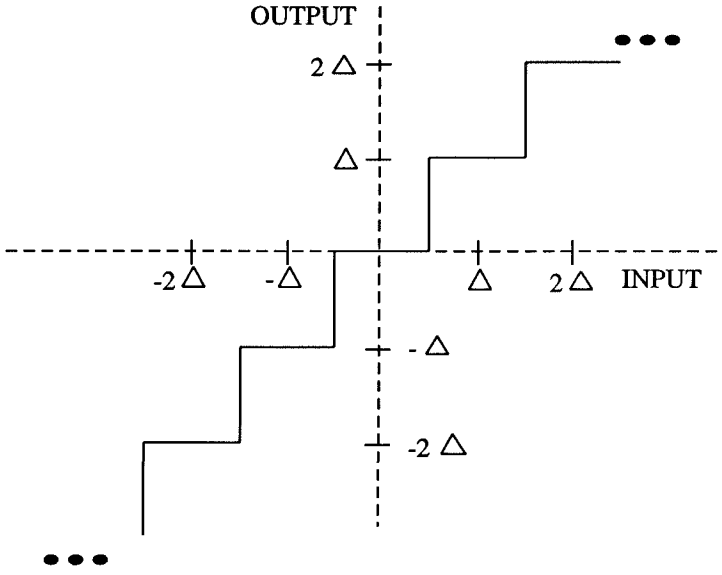


Fig. 3: Conceptual waveform generator model

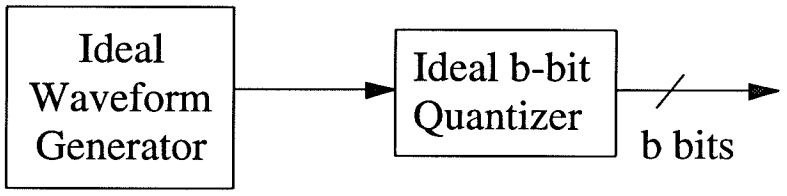


Fig. 4: Two-step waveform generator model

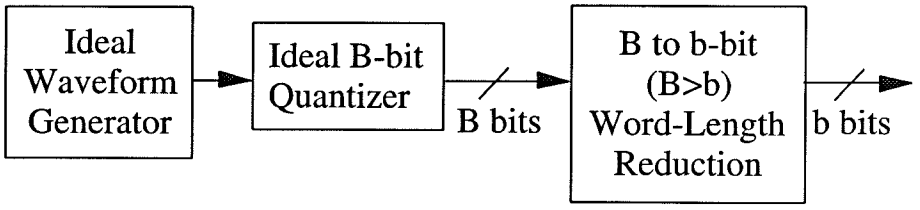


Fig. 5: Uniform dithered quantizer

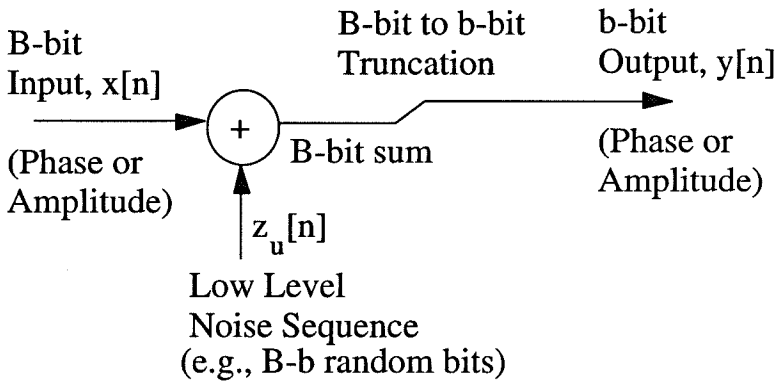


Fig. 6: Phase Quantization

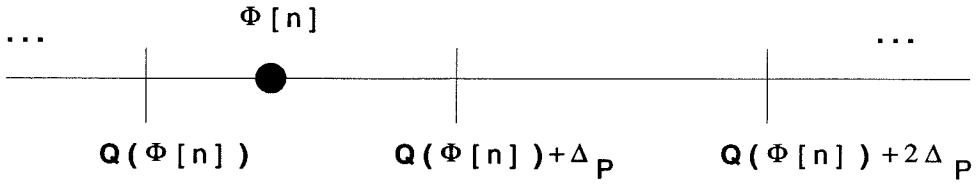


Fig. 7: First-Order Dithering: Quantization Noise Second Moment

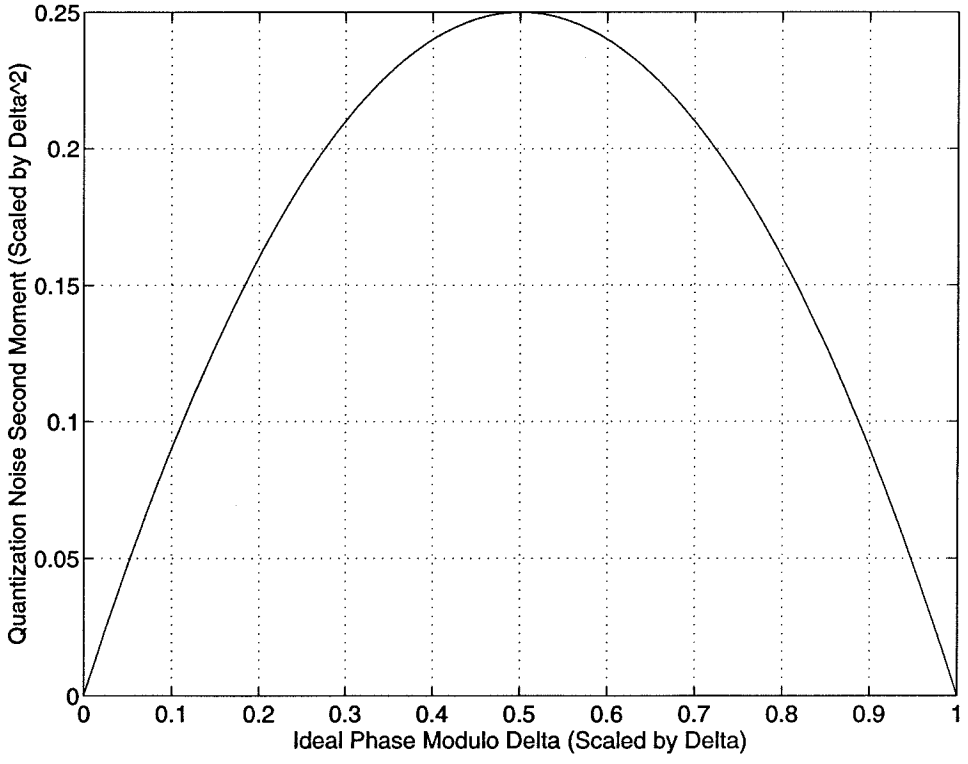


Fig. 8: Second-Order Dithering: Quantization Noise Third Moment

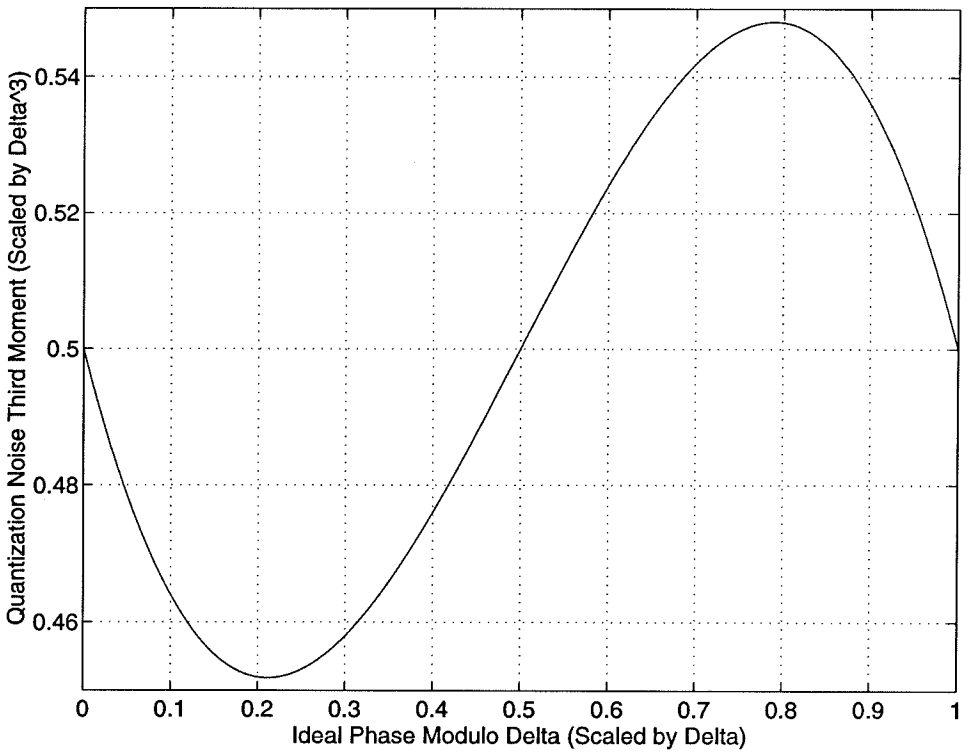


Fig. 9: Third-Order Dithering: Quantization Noise Fourth Moment

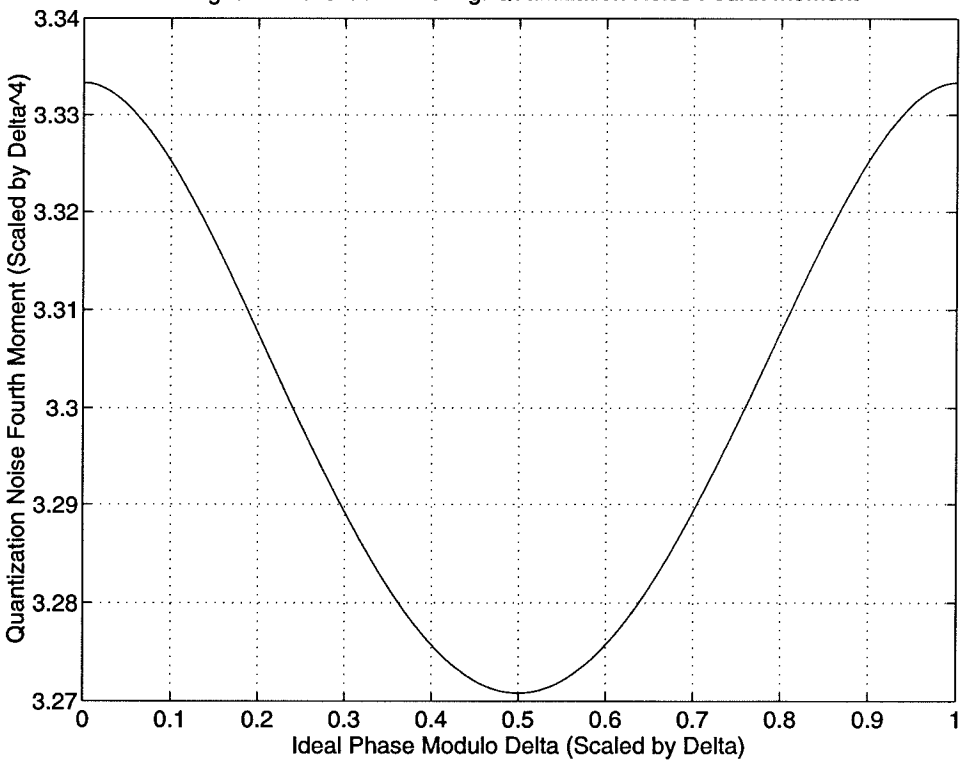


Fig. 10: Fourth-Order Dithering: Quantization Noise Fifth Moment

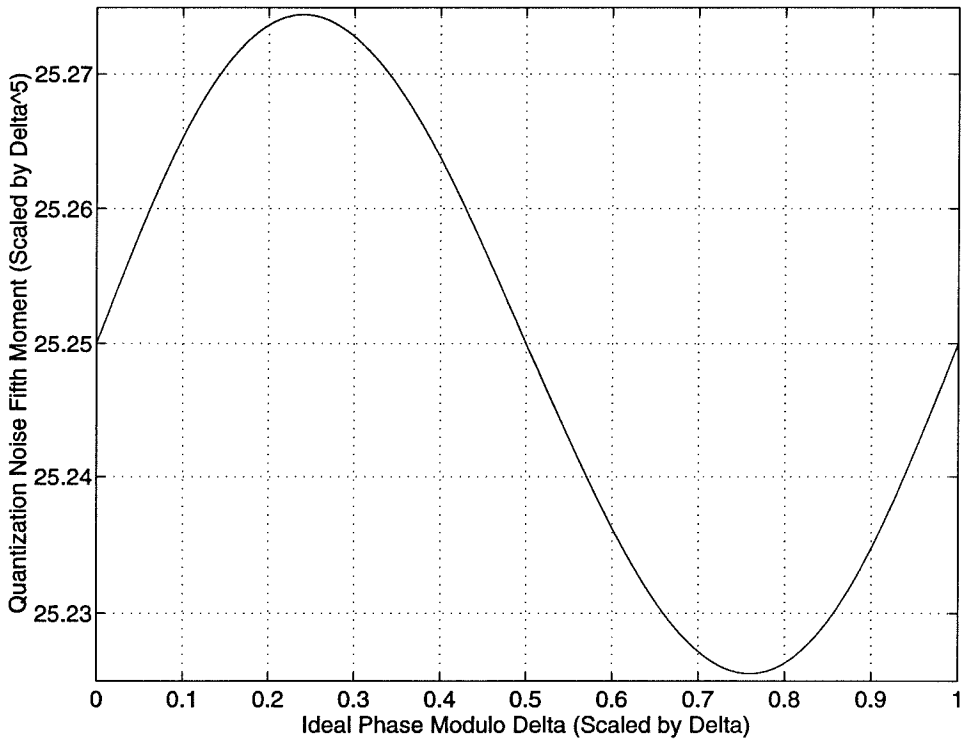


Fig. 11: Fifth-Order Dithering: Quantization Noise Sixth Moment

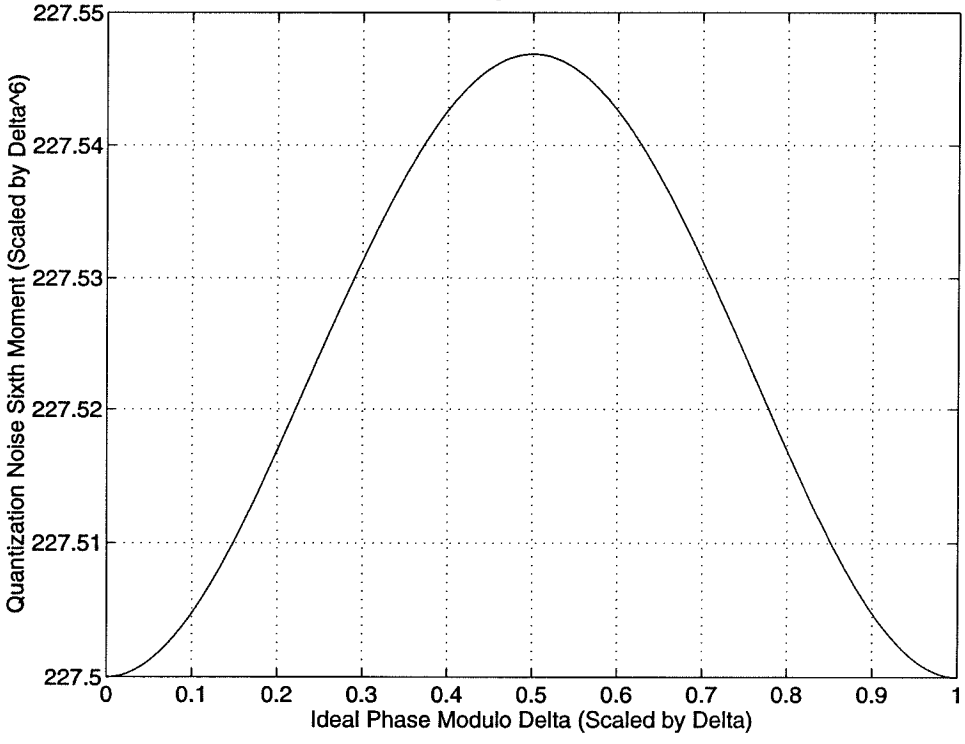


Fig. 12: Worst-Case Spur Bounds vs. Dither Order,M, and Phase Bits

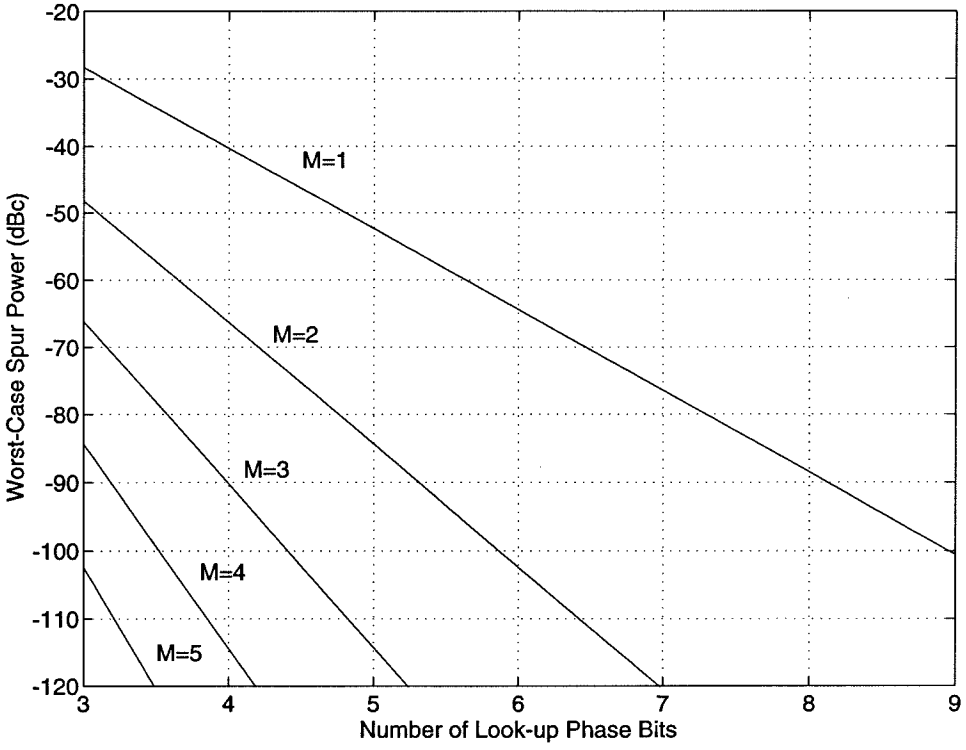


Fig. 13: Power spectrum of 8 sample/cycle sine wave without dithering (8 bit amplitude quantization)

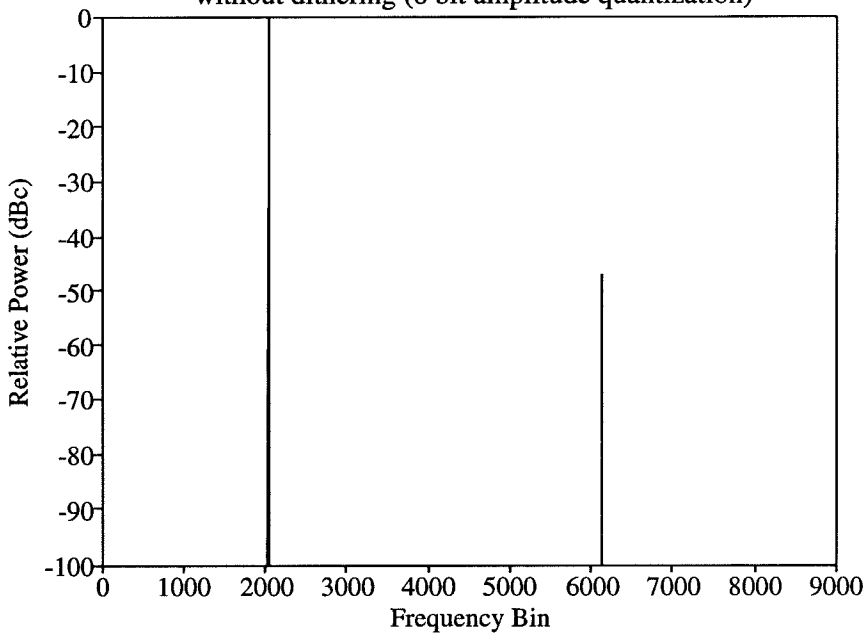


Fig. 14: Power spectrum of 8 sample/cycle sine wave with amplitude dithering (8 bit amplitude quantization)

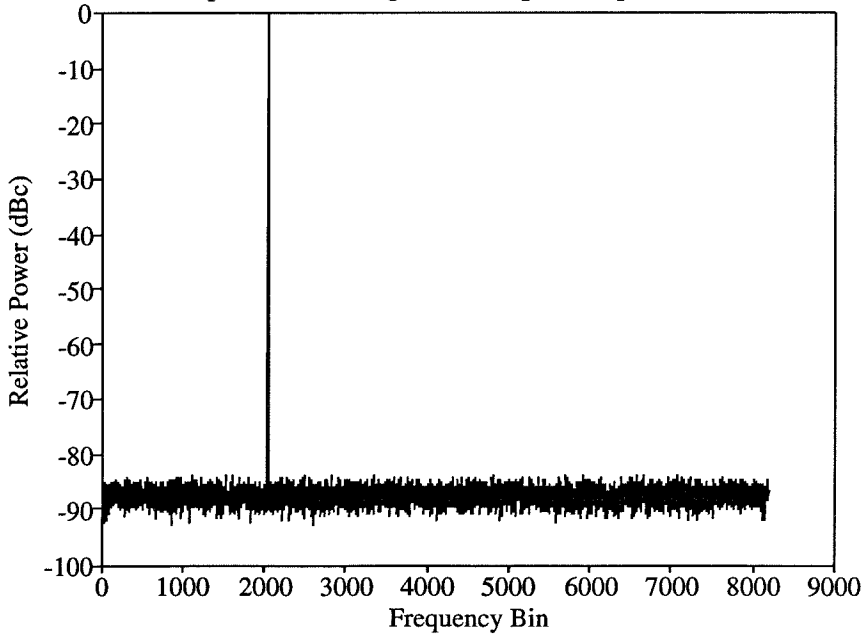


Fig. 15: Power spectrum of 5-bit phase-truncated sine wave without phase dithering (high-precision amplitude)

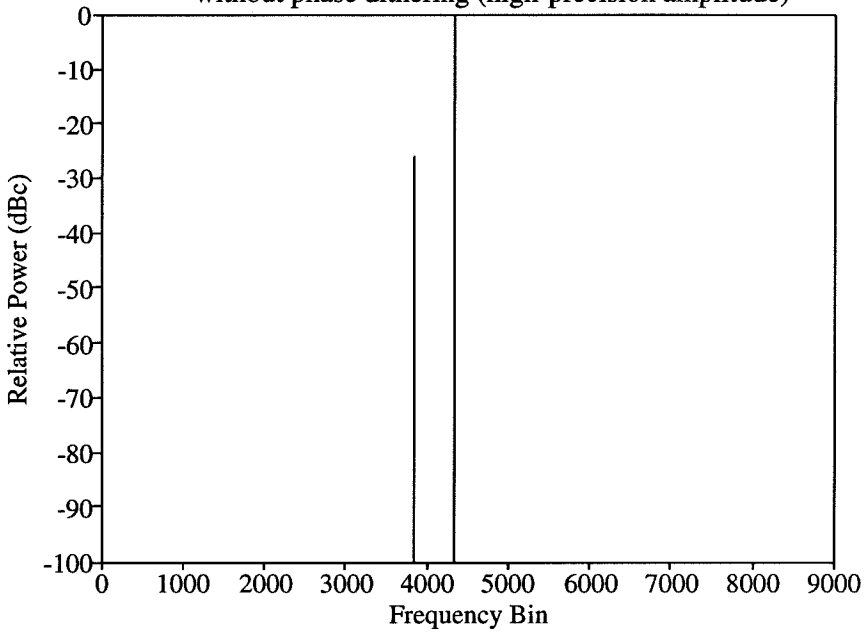


Fig. 16: Power spectrum of 5-bit phase-truncated sine wave with first-order phase dithering (high-precision amplitude)

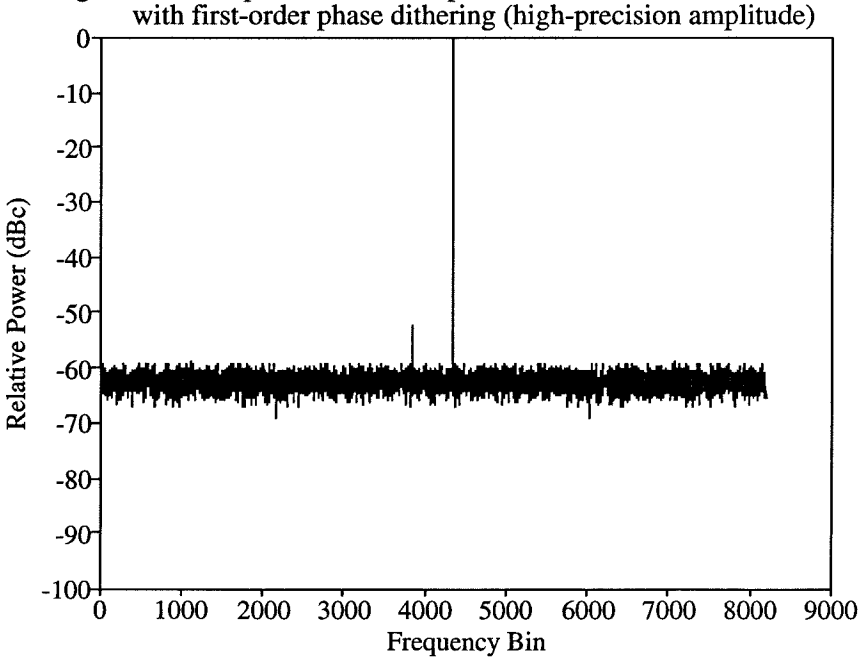


Fig. 17: Power spectrum of 5-bit phase-truncated sine wave with second-order phase dithering (high-precision amplitude)

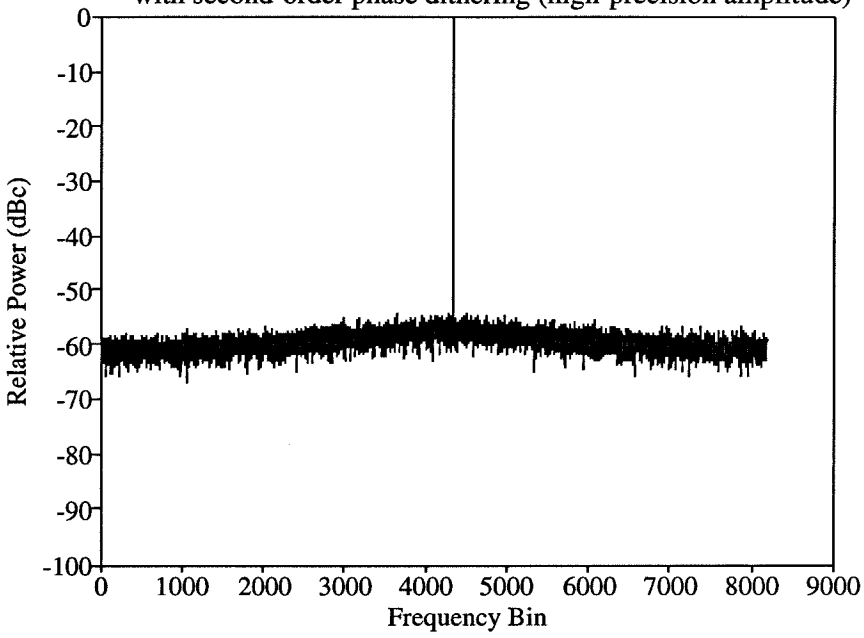


Fig. 18: Worst-case power spectrum of sinusoid with first-order phase dithering and amplitude dithering (8 bits each).

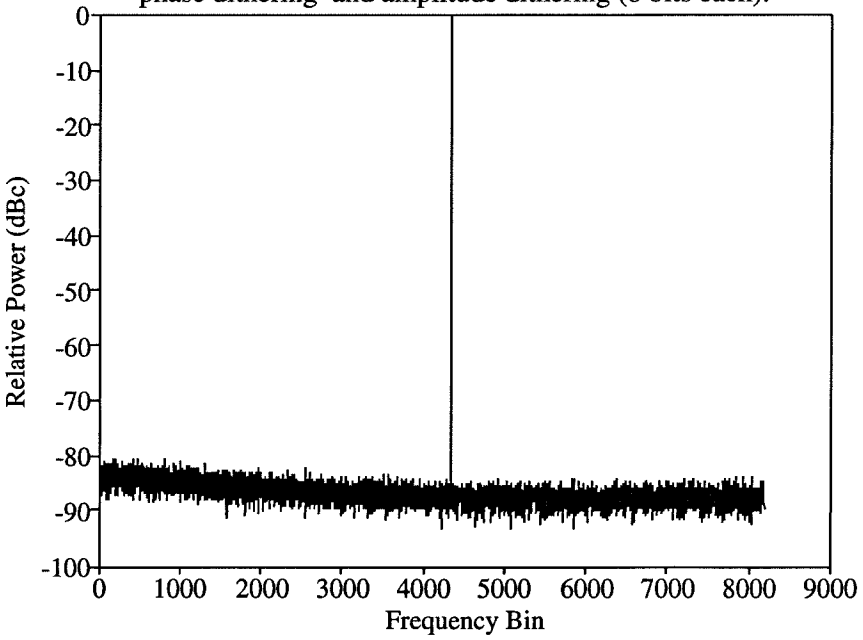


Fig. 19: 2nd-Order Dither, 4 Phase Bits

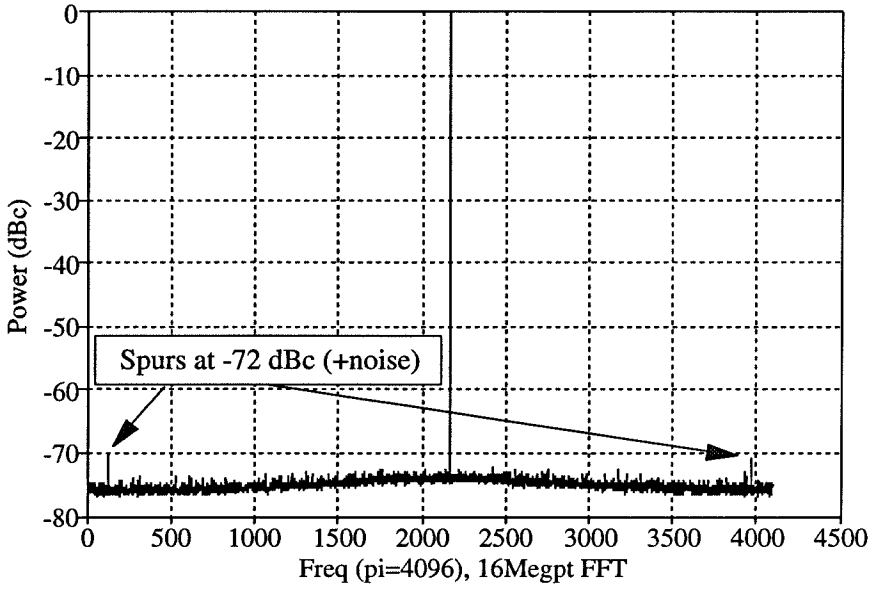


Fig. 20: Block diagram of first-order phase dithering system

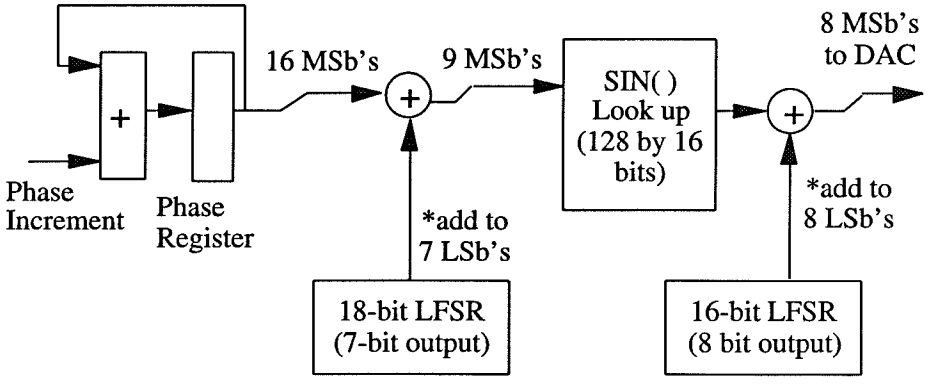
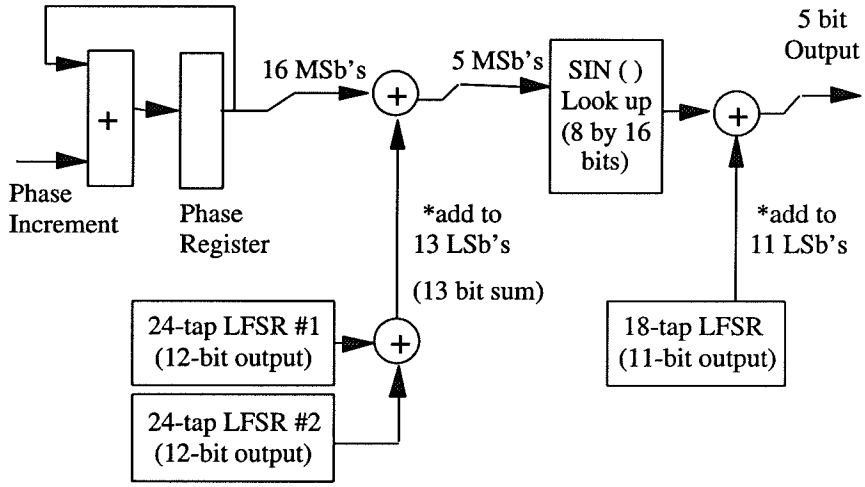


Fig. 21: 2nd-Order Phase Dither System



Chapter 3

The Oversampled Data Converter Problem and a Synthetic Approach to its Solution

3.1 Introduction and Motivation

Data converters are important because they act as the conduit between digital signal processors and their environment. Digital signal processors, in their theoretical essence, operate on signals that assume values from a discrete set of amplitudes and are of interest only at discrete time instants. However, the analog world consists of signals that generally are described as continuous amplitude and continuous time. Data converters are devices that transform digital information into analog information, and vice versa.

Linearity is one of the major aspects of data converters. Applications ranging from Direct Digital Synthesizers (DDS) to Compact Disc^{TM} recording/playing systems often require highly linear data converters. It is important to precisely define the context in which linearity is considered. By definition, data converters are "non-linear" devices because they contain quantizers [1]. To consider this further, we shall concentrate on the digital-to-analog converter (DAC.) The analog portion of any DAC can be modeled as an ideal quantizer followed

by a mapping function. The mapping function reflects the non-discrete macroscopic nature of all analog systems. For example, in a DAC that outputs voltage, when the digital input is the integer M, the output of the ideal quantizer is $M\Delta$ volts, where Δ is the nominal step size. The mapping function makes the overall DAC output $M\Delta + \epsilon_M$, where ϵ_M is the error associated with the M^{th} output level and in general is different for each M.

In a static model, the mapping function of a B-bit DAC can be described as a polynomial of degree $2^B - 1$ whose coefficients are determined from the 2^B analog levels corresponding to each quantizer state. This is the context in which linearity is considered: a DAC is declared "linear" when its mapping function is a first-order polynomial. While "linear" is an admittedly inappropriate expression, it is ubiquitous in the commercial and technical literature, and will be used here for historical purposes. For a multi-bit DAC ($B > 1$) it is impossible in practice to make the converter strictly linear and time-static [2] (for a time-varying strictly linear converter, see [5]). This would require making the non-linearity polynomial consist of a constant plus a first-order term. Multi-bit DACs are often referred to as "highly linear" when the magnitudes of the higher-order coefficients in the non-linearity polynomial are sufficiently small. This generally requires the use of expensive analog circuitry or trimming techniques. In contrast, a single-bit DAC is automatically linear using the static model. The single-bit analog-to-digital converter (ADC) may also be associated with a linear mapping function that precedes an ideal quantizer since there is only one threshold that the input is compared to when making a binary output decision. In general, it is better to use a dynamic model instead of a static model when describing the mapping function of an ADC or a DAC [4]. A dynamic model incorporates transient switching behavior of the analog output which undoubtedly has memory of previous states. The dynamic model greatly complicates analysis, however, and is not used in this analysis.

Conventional multi-bit data converters do not offer high linearity at high clock rates. For example, there are presently no conventional converters offering more than 14 bits of linearity at rates higher than 10 MHz [3]. Oversampled, coarse-resolution data converters offer high linearity and are of recent interest due to advances in VLSI technology [11]. This type of

data converter is attractive because it moves the burden of circuit complexity from the analog domain to the digital domain, resulting in a less expensive system suited to fabrication using common commercial digital VLSI techniques. These systems are "oversampled" because the digital clock rate exceeds the minimum rate required for sampling without aliasing. The *oversampling ratio*, R, is defined to be the ratio between the system sample rate and the Nyquist rate of the desired analog signal. "Coarse resolution" indicates that there are only a few quantization levels in the digital portion of the data converter. Coarse resolution systems permit tight control over linearity. This is because it is difficult to maintain each quantization level accurately with respect to other levels as the number of levels increases. While using only a few quantization levels facilitates good linearity, it also leads to a large amount of quantization noise. But if the bandwidth of the desired analog signal is smaller than the digital clock rate it is possible to place much of the quantization noise into the frequency regions unoccupied by the desired analog signal [11]. Subsequent frequency-selective filtering can then reject much of the quantization noise without degrading the desired signal. Oversampled data converters are often referred to as "noise shaping" systems. Approaches consistent with this appellation have been studied before [6, 7, 8]. Delta-Sigma ($\Delta\ \Sigma$) modulators are arguably the most popular of these approaches. The "noise shaping" approach is not taken here because it is fraught with the peril of unknowingly designing a system that is unstable [11]. Most $\Delta\Sigma$ modulators also require large oversampling ratios in order to provide acceptable noise performance. The parallel $\Delta\Sigma$ modulator ($\Pi\Delta\Sigma$) proposed by Galton [9] is a notable exception that does not require an oversampled input signal. However, each of the M parallel branches in the $\Pi\Delta\Sigma$ modulator heuristically can be viewed as a $\Delta\Sigma$ modulator system with oversampling ratio M. The burden of having a large oversampling ratio for satisfactory noise performance in conventional $\Delta\Sigma$ modulators is transformed into the burden of having a large number of parallel branches in the $\Pi\Delta\Sigma$ modulator.

The work in the second part of this thesis departs from the standard approach to the design of oversampled data converters. The standard approach (see [11] for many examples) typically presents the architecture first and then proceeds with an analysis of noise perfor-

mance. More recent alternatives to conventional $\Delta\Sigma$ architectures [12,13], while effective, often lack explicit theoretical motivation, and one naturally wonders if a given architecture is optimal in any sense. This part of the thesis presents *metric-based* approaches to the design of oversampled data converters. In this chapter and the next, noise performance metrics are presented and architectures that minimize these metrics are derived. This chapter next presents a formal description of the oversampled data converter problem. This problem is considered for the remainder of the thesis. This chapter then motivates, presents and analyzes a novel synthetic approach to the design of oversampled data converters.

3.2 The Oversampled Data Converter Problem

The oversampled data converter problem can be defined by the following question: given a large number, N, of samples of a discrete-time signal that is oversampled by a factor of R, what is the best discrete-time output sequence of length N that can be chosen whose samples all belong to a given, coarsely-quantized (and in the limiting case, binary) set? "Best" is meant in the sense that the quantization noise power that resides in the frequency region occupied by the desired signal is minimized over all legal output sequences of length N. This metric can be viewed as a classic mean-squared-error minimization of a filtered error and is a consequence of the assumption that an ideal frequency selective filter follows the data converter. This ideal filter passes the desired signal and rejects any signals that lie outside the desired frequency region. While no practical filters will be ideal, this assumption permits an analysis that concentrates on the limitations of the oversampled data converter and not the filtering operation.

There are several important problems related to the oversampled data converter problem described above. One related problem is the *a priori* computation of the amount of distortion that remains "in-band" as a function of the input signal, the oversampling ratio, the complexity of the system, and the nature of the coarsely-quantized output set. Another related problem is the scaling of the output set of data points. For example, given that

there are just two output sample values, $+c$ and $-c$, how should the constant c be chosen to minimize the amount of in-band quantization noise power?

3.3 The Synthesis Approach

A synthesis approach to the generation of a quantized sequence based on an oversampled input sequence is now presented. The generation algorithm is chosen to minimize a metric that measures the amount of error power that resides in the bandwidth occupied by the desired signal. The system that implements this generation algorithm is called the *Synthesis architecture.* It will be shown that the first-order $\Delta\Sigma$ modulator is a special case of the Synthesis architecture when the input is constant. The Synthesis architecture will also be shown to theoretically have superior in-band noise performance over the first-order $\Delta\Sigma$ modulator for finite oversampling ratios, and superior noise performance over conventional $\Delta\Sigma$ converters of arbitrary order when the oversampling ratio is less than 2.862.

Recent research in the field of Direct Digital Synthesis (DDS) has shown the need for high-linearity, high-speed digital-to-analog converters (DACs) that do not necessarily have high resolution [10]. In addition, such systems often generate signals that span large bandwidths relative to state-of-the-art digital clock rates. Therefore, the use of $\Delta\Sigma$ modulators [11] is often ruled out due to the low oversampling ratios that can be accommodated. The work presented here treats quantizers with an arbitrary number of bits, so the single-bit converter is a special case. In addition, this work takes into account potentially low oversampling ratios.

3.4 A Performance Metric and the Synthesis Architecture

It is desired to generate a quantized sequence, $y[n]$, based on an input sequence, $x[n]$. The values of $y[n]$ are to be chosen from an arbitrary quantized set. In the binary case, $y[n]$ is either

$+c$ or $-c$, where c is a constant. The output error sequence is defined as $e[n] = y[n] - x[n]$. Our goal is to select the output sequence that minimizes the error power in the frequency region $|\omega| < \pi/R$, where R is the oversampling ratio. The approach presented here *sequentially* chooses the values of $y[n]$ that minimize the following time-dependent performance metric:

$$\xi_n = \int_{-\pi/R}^{\pi/R} |E_n(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{3.1}$$

where

$$E_n(z) = \sum_{k=n-M+1}^{n} e[k] z^{-k} \tag{3.2}$$

is the windowed z-transform of the output error sequence. The metric in Equation 3.1 measures the in-band noise energy of the windowed error sequence by integrating the windowed error spectrum over the frequency region of interest. The metric is time-dependent because the windowed z-transform in Equation 3.2 is time-dependent. The integer constant M represents the memory of the system and limits the number of previous error samples that directly affect the metric. The aspect of finite memory ($M < \infty$) facilitates system realization and is largely ignored in $\Delta\Sigma$ modulators due to the ubiquity of simple integrators. While this finite memory approach was considered by Spang and Schultheiss [14], their development was analytic and required that the number of quantizer levels increase with the amount of memory, M. In this analysis, the number of quantization levels and the amount of memory are uncoupled quantities.

The following theorem presents the optimum generation algorithm that minimizes the distortion metric in Equation 3.1:

Theorem 3.1. *The optimum generation algorithm for the performance metric in Equation 3.1 is:*

$$y[n] = \mathrm{Q}\left(x[n] - \sum_{k=1}^{M-1} a_k e[n-k] \right), \tag{3.3}$$

where

$$a_k = \mathrm{sinc}(\frac{k}{R}) = \frac{\sin(\frac{\pi k}{R})}{\frac{\pi k}{R}} \tag{3.4}$$

and $Q(x)$ represents the legal quantized value closest to x.

Proof: At time instant n, a decision has to be made to make $y[n]$ equal to one of its legal quantized values. The value that minimizes the time-dependent metric in Equation 3.1 will be chosen. The windowed z-transform in Equation 3.2 can be written as the sum of a term independent of the choice at time instant n and a term dependent on the choice at time instant n:

$$E_n(z) = \hat{E}_n(z) + e[n]z^{-n}. \tag{3.5}$$

Taking the magnitude squared of the above equation and integrating to generate the metric in Equation 3.1 yields:

$$\xi_n = \hat{\xi}_n + \frac{1}{R}(y[n])^2 - \frac{2}{R}y[n]\left(x[n] - \sum_{k=1}^{M-1} a_k e[n-k]\right), \tag{3.6}$$

where $\hat{\xi}_n$ is independent of the choice of $y[n]$ at time instant n and the coefficients a_k are defined in Equation 3.4. Complete the square for the portion of ξ_n that is dependent on $y[n]$. To minimize ξ_n by our choice of $y[n]$, it is only necessary to minimize the portion of ξ_n dependent on $y[n]$. Therefore, it is equivalent to minimize

$$\left(y[n] - \left(x[n] - \sum_{k=1}^{M-1} a_k e[n-k]\right)\right)^2. \tag{3.7}$$

Therefore, Equation 3.3 is the optimum solution. ∎

The Synthesis architecture is shown in Fig. 1. It is important to note that the generation algorithm is dependent only on the present and past $M-1$ values of the input signal, $x[n]$. This makes sense since we select the $y[n]$ sequence term by term, suggesting that the future of $x[n]$ is unknown. The output sequence in Equation 3.3 is the quantization of the input signal plus a linear time-invariant filtered version of previous output errors. This bears a strong resemblance to the output of arbitrary $\Delta\Sigma$ modulators [11]. In fact, the following corollary shows that the first-order $\Delta\Sigma$ modulator is a special case of this synthesis approach.

Corollary 3.1. *The first-order $\Delta\Sigma$ modulator is a special case of the Synthesis architecture where the oversampling ratio, R, the memory, M, and their ratio, R/M, tend to infinity. This corresponds to the case of a nearly-constant input signal.*

Proof: As the ratio R/M tends to infinity, each of the coefficients defined in Equation 3.4 tend to unity since $0 \leq k < M$, and the output can be written as:

$$y[n] = \mathrm{Q}\left(x[n] - \sum_{k=1}^{M} e[n-k]\right). \tag{3.8}$$

In the standard $\Delta\Sigma$ literature, it is more common to express the output error sequence, $e[n]$, in terms of the error introduced by the data conversion. Let $\epsilon[n]$ be the quantization error introduced by the quantization operation, Q, in Equation 3.8. Then we may re-write the equation as:

$$y[n] = x[n] - \sum_{k=1}^{M} e[n-k] + \epsilon[n] = x[n] + e[n]. \tag{3.9}$$

Therefore as M tends to infinity:

$$\sum_{k=0}^{\infty} e[n-k] = \epsilon[n]. \tag{3.10}$$

It follows that under the asymptotic conditions described above, $e[n] = \epsilon[n] - \epsilon[n-1]$ which leads to the standard first-order $\Delta\Sigma$ result:

$$y[n] = x[n] + \epsilon[n] - \epsilon[n-1]. \tag{3.11}$$

As a check, the transfer function that the conversion error, $\epsilon[n]$, sees is $(1 - z^{-1})$, which has a spectral null at DC. ∎

3.5 In-band Noise Power

The in-band noise power is defined to be the amount of error power in the frequency region $|\omega| < \pi/R$. To calculate the in-band noise power, it is assumed that the conversion error, $\epsilon[n]$, made by the quantizer is a zero-mean, white, random variable uniformly distributed over one quantization interval. While this assumption is not exactly correct, in practice it can be a good first-order approximation for oversampled data converters [11], and it provides a tractable method of evaluating different architectures. It should be noted that the assumption that the quantizer can be replaced with an additive, white, uniformly-distributed

noise (AWUN) source is a significant departure from the treatment of quantizers in the first part of this thesis. In dithered digital sinusoid generators [10], the AWUN quantizer model is inappropriate because the quantizer input signals are relatively-pure periodic signals. These signals are corrupted by dither, but the magnitude of the dither is assumed to be small. In this chapter, the input signal is a band-limited Gaussian process and the validity of the AWUN quantizer model is investigated and shown to be satisfactory to a first-order approximation.

As in the $\Delta\Sigma$ literature, the difference equation relating the output error, $e[n]$, to the conversion error, $\epsilon[n]$, defines the *noise transfer filter,* $N(z)$, that the conversion error passes through to create the output error [11]. Replacing the quantizer by an additive error signal, $\epsilon[n]$, as in the proof of Corollary 3.1, we can write Equation 3.3 as:

$$y[n] = x[n] - \sum_{k=1}^{M-1} a_k e[n-k] + \epsilon[n] = x[n] + e[n],$$

since the output error, $e[n]$, is the difference of the input and output signals. Using Equation 3.4 and noting that $a_0 = 1$, the desired difference equation can be written as:

$$\epsilon[n] = \sum_{k=0}^{M-1} \mathrm{sinc}(\frac{k}{R}) e[n-k]. \tag{3.12}$$

Taking z-transforms we obtain:

$$\phi_e(z) = \frac{\phi_\epsilon(z)}{A(z)} = N(z)\phi_\epsilon(z), \tag{3.13}$$

where

$$A(z) = \sum_{k=0}^{M-1} \mathrm{sinc}(\frac{k}{R}) z^{-k}, \tag{3.14}$$

and $\phi_e(z)$ and $\phi_\epsilon(z)$ are the z-transforms of the output error and the conversion error, respectively.

The transfer function $A(z)$ can be viewed as a rectangularly-windowed FIR approximation to the IIR transfer function whose frequency response is R for $|\omega| < \pi/R$ and zero elsewhere. The approximation improves at frequencies away from $\pm\pi/R$ as the memory, M, increases. Therefore, over the frequency range, $|\omega| < \pi/R$, the magnitude of the noise transfer filter

response, $|N(e^{j\omega})|$, approaches $\frac{1}{R}$ as the memory increases. As the memory, M, becomes large, it follows that the in-band noise power is:

$$N_{IB} \rightarrow \sigma_\epsilon^2 \int_{-\pi/R}^{\pi/R} |N(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \frac{\sigma_\epsilon^2}{R^3}, \quad M \rightarrow \infty, \tag{3.15}$$

where σ_ϵ^2 is the variance of the conversion error.

The dependence of the in-band noise power on the inverse third power of the oversampling ratio is the same here as it is for first-order $\Delta\Sigma$ modulators [11]. However, an important difference is that the above relationship is valid for arbitrary oversampling ratios using this synthesis approach, while the $\Delta\Sigma$ result is valid only for large oversampling ratios. The next theorem elaborates on this result.

Theorem 3.2. *As the memory becomes infinite, the Synthesis architecture provides superior in-band noise performance over the first-order $\Delta\Sigma$ modulator for non-constant input signals.*

Proof: When the input signal is constant, the oversampling ratio is infinite, and Equation 3.15 shows that the in-band noise power is zero for the Synthesis architecture. From the proof of Corollary 3.1, the spectral null at DC in the noise transfer function for the first-order $\Delta\Sigma$ modulator shows that the in-band noise power is also zero.

It is well-known that the noise transfer function for the first-order $\Delta\Sigma$ modulator is $N(e^{j\omega}) = (1 - e^{-j\omega})$. Therefore the in-band noise power for the first-order $\Delta\Sigma$ modulator is:

$$N_{IB}|_{\Delta\Sigma 1} = \sigma_\epsilon^2 \int_0^{\pi/R} |1 - e^{-j\omega}|^2 \frac{d\omega}{\pi} = \frac{2}{R}\sigma_\epsilon^2 \left(1 - \mathrm{sinc}(\frac{1}{R})\right). \tag{3.16}$$

The ratio of the in-band noise power of the first-order $\Delta\Sigma$ modulator and the in-band noise power of the Synthesis architecture is a function of R and is found by dividing Equation 3.16 by the result of Equation 3.15:

$$f(R) = \frac{N_{IB}|_{\Delta\Sigma 1}}{N_{IB}} = 2R^2(1 - \mathrm{sinc}(\frac{1}{R})). \tag{3.17}$$

The ratio is greater than one if and only if $\sin(\pi/R) < \frac{\pi}{R}(1 - \frac{1}{2R^2})$. Using the product series expansion for the $\sin(\theta)$ function, the inequality

$$\sin(\frac{\pi}{R}) = \frac{\pi}{R}\prod_{k=1}^{\infty}(1 - \frac{1}{R^2k^2}) = \frac{\pi}{R}(1 - \frac{1}{R^2})\prod_{k=2}^{\infty}(1 - \frac{1}{R^2k^2}) < \frac{\pi}{R}(1 - \frac{1}{2R^2}) \tag{3.18}$$

is true because, after the obvious cancellation, $(1 - \frac{1}{R^2}) < (1 - \frac{1}{2R^2})$ and $|1 - \frac{1}{R^2k^2}| < 1$ are true for $R > 1$ and $k \geq 2$.

∎

It is natural to wonder how well the Synthesis architecture compares to conventional $\Delta\Sigma$ modulators of arbitrary order. In-band noise power calculations for an L^{th} order $\Delta\Sigma$ modulator can be made for arbitrary R by replacing the $|1-e^{-j\omega}|^2$ integrand in Equation 3.16 by $|1-e^{-j\omega}|^{2L}$. The solution to this integral is easily found. The L^{th} order $\Delta\Sigma$ modulator has in-band noise performance like $O(R^{-2L-1})$ as the oversampling ratio, R, becomes large [11]. Therefore, high-order $\Delta\Sigma$ modulators will have lower in-band noise performance than the Synthesis architecture for large oversampling ratios. However, they will perform more poorly for lower oversampling ratios. It is of interest to see what the critical oversampling ratio is in order for the L^{th} order $\Delta\Sigma$ modulator to have the same in-band noise performance as the Synthesis architecture. This can be viewed as the minimum oversampling ratio at which the Synthesis architecture is outperformed. These values have been computed as a function of the order, L, and are presented in Fig. 2. At oversampling ratios below approximately 2.862, no $\Delta\Sigma$ modulator of any order outperforms the Synthesis architecture. Therefore, this simple analysis suggests the Synthesis architecture can outperform $\Delta\Sigma$ modulators of arbitrary order in high-bandwidth applications requiring a highly linear data converter.

3.6 Simulations

Fig. 3 shows the theoretical in-band noise performance of the Synthesis architecture and the first-order $\Delta\Sigma$ modulator for oversampling ratios between 1 and 10. These low oversampling ratios are of interest based on the results at the end of the last section. The in-band power is scaled by the variance of the conversion error, σ_ϵ^2. As proved in Theorem 3.2, the Synthesis architecture theoretically performs better than the first-order $\Delta\Sigma$ modulator under the white quantization noise assumption.

Simulations were performed for oversampling ratios equal to $2, 4$ and 8. The input signal

for each 128K-point simulation was a stationary Gaussian process with zero-mean, unit variance and frequency support constrained to the region $|\omega| < \pi/R$ by an elliptic filter of appropriate order to have at least 0.3 dB passband ripple and 70 dB stopband attenuation. The same filter was used on the output error sequences before evaluating the experimental in-band noise power. Single-bit converters with output values $\pm c = \pm 2.5$ were used in all simulations. The converter output magnitude, $|c|$, was chosen so that the unit-variance Gaussian would overload, or saturate, the converter infrequently.

The circles (o) and asterisks (*) in Fig. 3 are the experimental outcomes for the $\Delta\Sigma$ modulator and the Synthesis architecture with memory of $M = 100$, respectively. While the experimental results indicate that for these oversampling ratios the Synthesis architecture outperforms the first-order $\Delta\Sigma$ modulator as predicted by theory, the margin by which it does so and the actual values of the in-band noise power depart from what is expected. These variations are explained by the fact that for each simulation the spectrum of the conversion error process, $\epsilon[n]$, was not exactly white. While conversion error variance was very close to the predicted values of $\sigma_\epsilon^2 = c^2/3$, the conversion error spectrum was slightly high-pass for the $\Delta\Sigma$ modulator and resulted in lower in-band noise power. Conversely, the conversion error spectrum was slightly low-pass for the Synthesis architecture resulting in greater in-band noise power. The theoretical expressions derived in Section 3.5 assumed white conversion noise, which is valid to a first-order approximation judging from the experimental outcomes in Fig. 3 and separate simulations that studied the conversion error specifically.

The simulations also verified predictions about the nature of the noise transfer function for the Synthesis architecture. Spectra of the output error were obtained and found to have approximately constant response over the passband frequencies. The relationship between in-band noise power and memory was investigated and results are presented in Fig. 4 for oversampling ratios of 4 and 8. It was shown in Section 3.5 that the magnitude of the noise transfer function in the passband tended to $\frac{1}{R}$, where R is the oversampling ratio, as the memory, M became large. Fig. 4 suggests that this will be true when the ratio of the memory to the oversampling ratio exceeds unity, or when $M > R$. This has important implications

for the complexity of Synthesis architectures. The number of delays and multipliers required in a system that approaches the asymptotic performance described in Equation 3.15 may be small when the oversampling ratio is not large.

3.7 Conclusions

The Synthesis architecture is the optimal solution to an oversampled data conversion problem based on the minimization of the metric in Equation 3.1. While the Synthesis architecture was theoretically and experimentally shown to have better in-band noise performance than the first-order $\Delta\Sigma$ modulator, the Synthesis architecture has greater complexity. The added complexity may be worthwhile, however, in systems with low oversampling ratios. In addition, experimental results have shown that in such systems the necessary complexity to approach asymptotic performance is not large.

The limiting aspect of the analytical model presented here was the behavior of the conversion error. This was the primary reason for discrepancy between the theoretical predictions for in-band noise power and experimental observations. Future work needs to address more complex models for the conversion error spectra as a function of the input signal.

References

1. R.M. Gray, "Quantization Noise Spectra," *IEEE Transactions on Information Theory,* vol. IT-36, pp. 1220-1244, November 1990.

2. *Analog-Digital Conversion Handbook,* edited by D. Sheingold, Prentice-Hall, Englewood Cliffs, NJ, 1986.

3. *Electronic Design Magazine,* page 111, September 16, 1993.

4. G.R. Spalding, R.L. Geiger, "Digital Correction for Improved Spectral Response in Signal Generation Systems," *IEEE Proc. ISCAS'93,* pp. 132-135, May 1993.

5. P. Carboni, I. Galton, "A Rigorous Error Analysis of D/A Conversion with Dynamic Element Matching," submitted to *IEEE Transactions on Circuits and Systems,* October 1993.

6. S.K. Tewksbury, R.W. Hallock, "Oversampled, Linear Predictive and Noise-Shaping Coders of Order $N > 1$," *IEEE Transactions on Circuits and Systems,* vol. CAS-25, pp. 436-447, July 1978.

7. Y. Matsuya, K. Uchimura, A. Iwata, T. Kobayashi, M. Ishikawa, T. Yoshitome, "A 16-Bit Oversampling A-to-D Conversion Technology Using Triple-Integration Noise Shaping," *IEEE Journal of Solid-State Circuits,* vol. SC-22, pp. 921-929, December 1987.

8. K. Uchimura, T. Hayashi, T. Kimura, A. Iwata, "Oversampling A-to-D and D-to-A Converters with Multistage Noise Shaping Modulators," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-36, pp. 1899-1905, December 1988.

9. I. Galton, "An Analysis of Quantization Noise in $\Delta\Sigma$ Modulation and its Application to Parallel $\Delta\Sigma$ Modulation," Ph.D. dissertation, Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, 1992.

10. M. Flanagan, G. Zimmerman, "Spur-Reduced Digital Sinusoid Synthesis," accepted for publication,*IEEE Transactions on Communications.*

11. *Oversampled Data Converters: Theory, Design and Simulation,* edited by J. Candy, G. Temes, IEEE Press, New York, 1992.

12. K. Chao, S. Nadeem, W. Lee, C. Sodini, "A Higher Order Topology for Interpolative Modulators for Oversampling A/D Converters," *IEEE Transactions on Circuits and Systems,* vol. CAS-37, pp. 309-318, March 1990.

13. P. Ferguson, Jr., A. Ganesan, R. Adams, "One Bit Higher Order Sigma-Delta A/D Converters," *IEEE Proc. ISCAS'90,* pp. 890-893, May 1990.

14. H. Spang, P. Schultheiss, "Reduction of Quantizing Noise by Use of Feedback," *IRE Transactions on Communications Systems,* pp. 373-380, December 1962.

Figures

1. The Synthesis architecture
2. Minimum R to outperform the Synthesis architecture
3. In-band noise power vs. R, theory and experiment
4. In-band noise versus memory, R=4,8

Fig. 1: The Synthesis Approach System Diagram

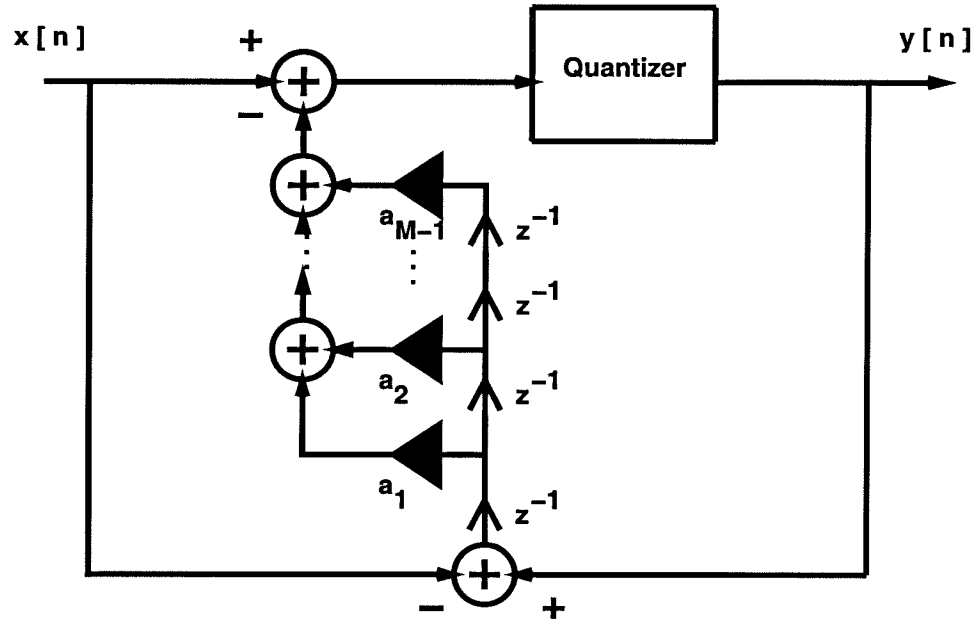


Fig. 2: Minimum R to Outperform Synthesis Architecture

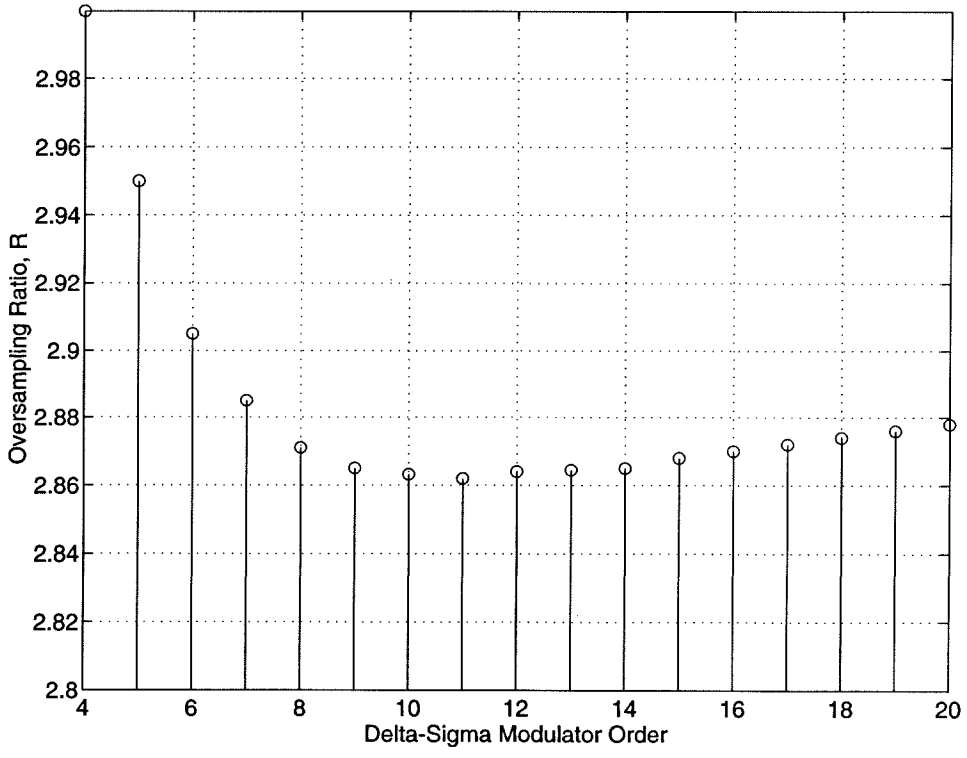


Fig. 3: In-band Noise Power vs. R, Theory and Experiment

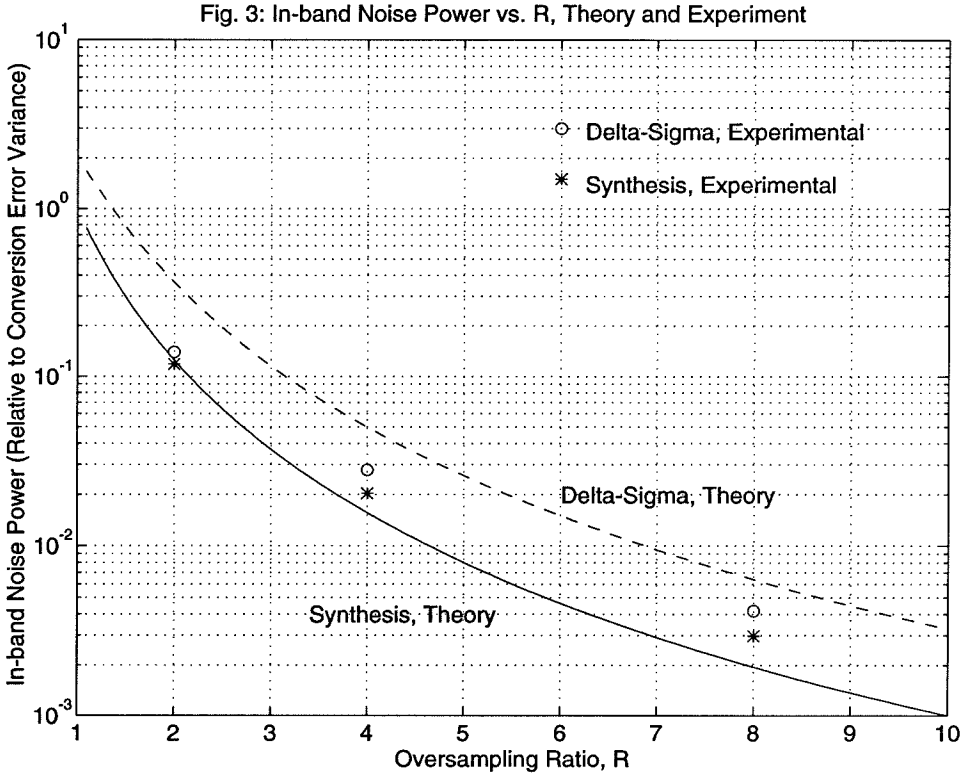
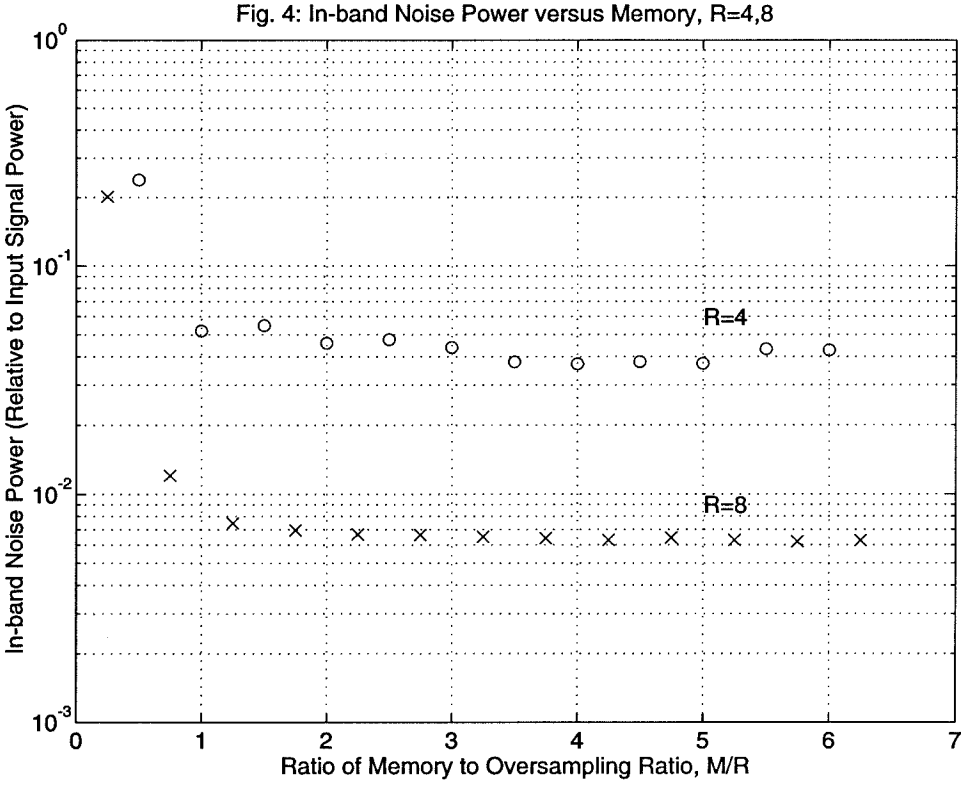


Fig. 4: In-band Noise Power versus Memory, R=4,8



Chapter 4

Eigenmodulators: A Vector Quantization Approach to Oversampled Data Conversion

4.1 Motivation and Background

This chapter presents a metric-based vector quantization approach to the analysis and design of oversampled data converters. This approach differs from the synthesis approach in the previous chapter. The previous synthesis approach produced only a single output value based on a single input value and a block of past errors. The present approach considers the generation of blocks of output data based on blocks of input data. For example, the block of input data may be taken to be as large as the entire input data sequence. The metric-based analysis leads to a framework for viewing the data converter problem in light of related eigenvectors and eigenvalues. This framework motivates a new architecture for oversampled data converters that is based on a vector quantization system.

This vector quantization approach associates a real, Hermitian, positive-definite matrix with an appropriate power spectral distortion metric. The power spectral distortion metric reflects the amount of error power that resides in the frequency regions occupied by the

desired signal. The orthonormal eigenvectors of the Hermitian matrix are used to map the desired input signal and all possible output signals onto the space spanned by the eigenvectors. The input signal is analog in the Analog-to-Digital converter (ADC) and digital in the Digital-to-Analog converter (DAC). This framework applies to both DACs and ADCs. The mappings of the input and output signals onto the eigenspace generated by the eigenvectors provide a convenient means for evaluating the output distortion, and hence, a means for choosing output vectors based on input vectors. The *eigenmodulator* is an oversampled data converter that operates in this manner. While this framework can be used in scenarios where the frequency region occupied by the desired signal is away from DC, the important case of low-pass oversampled data converters will be considered exclusively. It will be shown that the eigenvectors are the discrete prolate spheroidal wave sequences described in detail by Slepian [1].

The transformation used in this chapter is well-known. Huang and Schultheiss [2] used a similar vector transformation to quantize a block of correlated Gaussian random variables. However, they then transmitted the quantized components of the transformed vectors over a digital channel. The framework in this chapter does not transmit these transformed vectors but instead uses them to decide which quantized vector should be chosen to minimize a distortion metric for a given input vector. The transformation has also been successfully used in the rate-distortion analysis [5] of Gaussian signals with squared-error distortion metrics. The rate-distortion analysis in this chapter is based on this work.

This chapter introduces a vector quantization problem that is asymptotically equivalent to the oversampled data converter problem described in the previous chapter. An important consequence of this equivalence is that the growing field of vector quantization theory [3] may then be applied to the analysis and design of oversampled data converters. To illustrate this, an application of vector quantization theory to the selection of the output levels in a single-bit data converter is considered in the next chapter.

The next section presents the power spectral distortion metric used for the rest of this chapter and the next chapter. The metric is presented in the context of a conventional signal

processing filtering problem. Then a linear data transformation is introduced that turns the distortion metric into a squared-error metric. The following section presents asymptotic properties of the data transformation as larger blocks of data are transformed. After a discussion of the eigenmodulator, the chapter concludes with a rate-distortion analysis of the oversampled data converter problem that is facilitated by the data transformation.

4.2 A Power Spectral Distortion Metric

Consider an oversampled data converter with a bounded real input sequence, $x[n]$, where $n = 0, 1, ..., N-1$. An input vector, $\mathbf{x}$, can be constructed where $\mathbf{x}^\dagger = [x[0]x[1]\cdots x[N-1]] \in \mathcal{R}^N$ and $\mathcal{R}$ is the set of real numbers. Vector transposition and conjugation are denoted by the dagger symbol, $\dagger$. It is desired to generate an N-sample output sequence based on the input sequence. We are constrained to choose the output vector from $\mathcal{Q}$, an arbitrarily-ordered, N-dimensional set of $|\mathcal{Q}|$ quantized vectors. The l^{th} vector in $\mathcal{Q}$ is denoted by $\mathbf{y}_l$ where $l = 0, 1, 2, ..., |\mathcal{Q}| - 1$, and has the form $\mathbf{y}_l{}^\dagger = [y_l[0]y_l[1]\cdots y_l[N-1]]$. It is often the case that $\mathcal{Q} = \mathcal{L}^N$ where $\mathcal{L}$ is a set of scalar output values. For example, in the single-bit case, $\mathcal{L}$ could be $\{+c, -c\}$, where c is some scaling constant. The error vector, $\mathbf{e}_l$, is defined to be the difference between the l^{th} output vector and the input vector, or $\mathbf{e}_l = \mathbf{y}_l - \mathbf{x}$. The n^{th} element of the error vector is $e_l[n] = y_l[n] - x[n]$.

It is desired to minimize the error power over the frequency interval, $\Omega \equiv (-\frac{\pi}{R}, \frac{\pi}{R})$, where R is the oversampling ratio. The distortion metric used in this problem is the average error power over Ω:

$$D_l = \frac{1}{N}\int_\Omega |E_l(e^{j\omega})|^2 \frac{d\omega}{2\pi}, \tag{4.1}$$

where $E_l(z)$ is the z-transform of the error sequence associated with the l^{th} output vector. The following lemma associates a Hermitian, positive-definite matrix with the error power metric in Equation 4.1.

Lemma 4.1. *The distortion metric in Equation 4.1 can be written as:*

$$D_l = \frac{1}{N}(\mathbf{y}_l - \mathbf{x})^\dagger \mathbf{A}(\mathbf{y}_l - \mathbf{x}), \tag{4.2}$$

where $\mathbf{y}_l$ *is the* l^{th} *output vector,* $\mathbf{x}$ *is the input vector, and the filter matrix* $\mathbf{A}$ *is a real Hermitian, positive-definite matrix defined by:*

$$\mathbf{A}_{n,m} = \int_\Omega \cos((n-m)\omega)\frac{d\omega}{2\pi} = \frac{\sin(\frac{\pi(n-m)}{R})}{\pi(n-m)} = \frac{1}{R}\mathrm{sinc}(\frac{(n-m)}{R}) \quad . \tag{4.3}$$

Proof: The z-transform of the error is defined as

$$E_l(z) = \sum_{n=0}^{N-1} e_l[n]z^{-n}.$$

Since $e_l[n] = y_l[n] - x[n]$, the z-transform can be written in vector notation as $E_l(z) = (\mathbf{y}_l - \mathbf{x})^\dagger \zeta(z)$, where $\zeta(z)$ is a column vector whose k^{th} element is z^{-k}. Therefore, the power metric in Equation 4.1 can be written as:

$$D_l = \mathrm{Re}(D_l) = \frac{1}{N}(\mathbf{y}_l - \mathbf{x})^\dagger \left(\int_\Omega \mathrm{Re}(\zeta(e^{j\omega})\zeta^\dagger(e^{j\omega}))\frac{d\omega}{2\pi} \right) (\mathbf{y}_l - \mathbf{x}),$$

which leads to Equations 4.2 and 4.3 and shows that $\mathbf{A}$ is real.

It remains to show that $\mathbf{A}$ is Hermitian and positive-definite. $\mathbf{A}$ is Hermitian by inspection of Equation 4.3 since $\mathbf{A}_{n,m} = \mathbf{A}_{m,n}$ due to the symmetry of the sinc function. $\mathbf{A}$ is positive-definite by the equality of Equations 4.1 and 4.2 since the former equation is strictly greater than zero whenever the error sequence is not identically zero.

∎

Since $\mathbf{A}$ is real, Hermitian and positive-definite, its eigenvectors are real and orthonormal and span the N-dimensional vector space [4]. Therefore, any N-dimensional vector, $\mathbf{z}$, can be written as a linear combination of the eigenvectors:

$$\mathbf{z} = \sum_{k=0}^{N-1} \frac{R^{1/2}}{\lambda_k^{1/2}} \hat{z}[k]\mathbf{v}_k, \tag{4.4}$$

where R is the oversampling ratio, $\mathbf{v}_k$ is the k^{th} eigenvector and λ_k the k^{th} eigenvalue satisfying the eigenvector equation:

$$\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k. \tag{4.5}$$

Let the eigenvectors be ordered by their eigenvalues, λ_k, where $1 \geq \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} > 0$. The eigenvalues are strictly positive because the matrix $\mathbf{A}$ is positive-definite. The eigenvalues in this problem are also less than or equal to unity as shown in Appendix II. The filter matrix may more generally be expressed as $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\dagger$, where $\Lambda = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$ is the diagonal matrix containing the eigenvalues and $\mathbf{V}$ is the unitary matrix ($\mathbf{V}^\dagger\mathbf{V} = \mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$) that diagonalizes the filter matrix. The eigenvector $\mathbf{v}_k$ is the kth column of $\mathbf{V}$.

The vector $\hat{\mathbf{z}}$ is the mapping of the input vector onto the space defined by the eigenvectors. For this reason, it is called the "eigenmapping". The eigenmapping is defined as:

$$\hat{\mathbf{z}} = R^{-\frac{1}{2}}\Lambda^{1/2}\mathbf{V}^\dagger\mathbf{z}. \tag{4.6}$$

Since the eigenvalues are non-zero and the matrix $\mathbf{V}$ is unitary, this linear transformation is non-singular and $\mathbf{z} = R^{1/2}\Lambda^{-1/2}\mathbf{V}\hat{\mathbf{z}}$. The uniqueness of the eigenmapping follows from the orthonormality of the eigenvectors.

These results lead to our first theorem which expresses the power metric in Equation 4.1 in terms of the eigenvalues of $\mathbf{A}$, and the eigenmappings of the input and output vectors.

Theorem 4.1. *The distortion metric in Equation 4.1 may be written in squared-error form:*

$$D_l = \frac{R}{N}\sum_{k=0}^{N-1}(\hat{x}[k] - \hat{y}_l[k])^2 = \frac{R}{N}(\hat{\mathbf{x}} - \hat{\mathbf{y}}_l)^\dagger(\hat{\mathbf{x}} - \hat{\mathbf{y}}_l), \tag{4.7}$$

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}_l$ are the eigenmappings of the input vector and the l^{th} output vector, respectively.

Proof: Using the fact that $\mathbf{A}$ can be diagonalized, we can write Equation 4.2 as:

$$D_l = \frac{1}{N}(\mathbf{x} - \mathbf{y}_l)^\dagger\mathbf{V}\Lambda\mathbf{V}^\dagger(\mathbf{x} - \mathbf{y}_l). \tag{4.8}$$

Multiplying both sides by $\frac{N}{R}$ and taking the square root of the eigenvalue matrix, Λ, gives:

$$\frac{N}{R}D_l = \frac{1}{R}(\mathbf{x} - \mathbf{y}_l)^\dagger\mathbf{V}\Lambda^{1/2}\Lambda^{1/2}\mathbf{V}^\dagger(\mathbf{x} - \mathbf{y}_l). \tag{4.9}$$

The desired result follows from the definition of the eigenmapping in Equation 4.6. ∎

The filter matrix, $\mathbf{A}$, defined in Equation 4.3, is the same matrix that was analyzed by Slepian [1] when he studied the spectra of index-limited sequences. Slepian showed that the eigenvectors and the eigenvalues of this matrix are related to the prolate spheroidal wave functions. Slepian also derived useful asymptotic properties, which we consider in the next section.

4.3 Asymptotic Properties

We now consider the results of the previous section as the block size, N, becomes large. We also consider the common case where the oversampling ratio, R, becomes large because many practical oversampled data converters employ large oversampling ratios. Systems with large oversampling ratios are studied in the next chapter. The asymptotic analysis in this section indicates that it is possible to consider only a certain number of terms in Equation 4.7 in order to choose an output sequence so that the resulting in-band distortion is below some level. This is useful for practical implementations as it reduces the dimension of the oversampled data converter problem. This result depends on analysis performed by Slepian regarding the eigenvalues of the filter matrix, $\mathbf{A}$. The next theorem presents the eigenvalue behavior for a fixed oversampling ratio. The proof is in Slepian's classic paper [1].

Theorem 4.2 (Slepian). *Let the oversampling ratio, R, be fixed. When the block size, N, tends to infinity and $k = \lfloor \frac{N}{R}(1-\epsilon) \rfloor$ with $0 < \epsilon < 1$, the eigenvalue λ_k tends to unity like*

$$1 - \lambda_k \sim Ce^{-LN} \tag{4.10}$$

where C and L are bounded positive constants. Similarly, for large block size and when $k = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$ with $0 < \epsilon < R-1$, the eigenvalue λ_k tends to zero like

$$\lambda_k \sim C'e^{-L'N} \tag{4.11}$$

where C' and L' are other bounded positive constants.

When ϵ is chosen to be slightly greater than zero it follows that approximately $\frac{N}{R}$ of the eigenvalues tend to unity and the remainder tend to zero. This indicates that the effective

dimension of the oversampled data converter problem is approximately the ratio of the block size to the oversampling ratio, $\frac{N}{R}$, instead of N. The following theorem modifies Theorem 4.2 to account for asymptotically large oversampling ratios. The proof appears in Appendix I of this chapter.

Theorem 4.3. *Let the block size N, the oversampling ratio R, and their ratio N/R all tend to infinity, and let $k = \lfloor \frac{N}{R}(1-\epsilon) \rfloor$ for fixed $0 < \epsilon < 1$. Then the eigenvalue λ_k tends to unity like*

$$1 - \lambda_k = O(e^{-\frac{\alpha N}{R}}), \tag{4.12}$$

where α is some positive constant. Similarly, when $k = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$ for fixed $0 < \epsilon < R-1$, the eigenvalue λ_k tends to zero like

$$\lambda_k = O(e^{-\frac{\alpha N}{R}}). \tag{4.13}$$

The next theorem expresses the distortion in the oversampled data converter problem as the sum of two distortion terms. The first term is the distortion due to a vector quantizer of dimension

$$N_{VQ} = \lfloor \frac{N}{R}(1+\epsilon) \rfloor \tag{4.14}$$

for some $\epsilon > 0$. The second distortion term is the residual distortion due to the dimensions not included in the vector quantization problem. It will be shown that the residual distortion is asymptotically small as the block size tends to infinity when all other parameters remain fixed. The dependence of the distortions on the index of the output vector, l, will now be dropped to simplify the notation.

Theorem 4.4. *The distortion metric originally presented in Equation 4.1 can be expressed as the sum of a vector quantization distortion and a residual distortion:*

$$D = D_{VQ} + D_{RESIDUAL},$$

where

$$D_{VQ} = \frac{R}{N} \sum_{k=0}^{N_{VQ}-1} (\hat{x}[k] - \hat{y}[k])^2 \quad . \tag{4.15}$$

Let the input vector and all output vectors have bounded average power. Then as the block size, N, tends to infinity, the oversampled data converter distortion tends to the vector quantization distortion, $D \rightarrow D_{VQ}$, because the residual distortion tends to zero like

$$D_{RESIDUAL} = O(e^{-L'N}),$$

for some positive constant L'.

Proof: The first claim of the proof follows trivially from Equation 4.7 in Theorem 4.1. To prove the second claim, we start by bounding the residual distortion:

$$D_{RESIDUAL} = \frac{R}{N} \sum_{k=N_{VQ}}^{N-1} \lambda_k \left(\frac{\hat{x}[k] - \hat{y}[k]}{\lambda_k^{1/2}} \right)^2 \leq \frac{\lambda_{N_{VQ}} R}{N} \sum_{k=0}^{N-1} \left(\frac{\hat{x}[k] - \hat{y}[k]}{\lambda_k^{1/2}} \right)^2 .$$

This follows from the facts that the eigenvalues are in descending order and that an upper bound is obtained when we increase the number of non-negative terms in the sum over k.

The final part of the proof uses a generalized Parseval's relation for orthonormal eigenvectors. Consider a vector $\mathbf{z}$ and its eigenmapping $\hat{\mathbf{z}}$. The sum of the squared components of the eigenmapping scaled by the appropriate eigenvalues can be written as:

$$\sum_{k=0}^{N-1} \left(\frac{\hat{z}[k]}{\lambda_k^{1/2}} \right)^2 = \hat{\mathbf{z}}^\dagger \Lambda^{-1} \hat{\mathbf{z}} = \frac{1}{R} \mathbf{z}^\dagger \mathbf{V} \Lambda^{1/2} \Lambda^{-1} \Lambda^{1/2} \mathbf{V}^\dagger \mathbf{z} = \frac{\mathbf{z}^\dagger \mathbf{z}}{R}.$$

This proves the generalized Parseval's property that the sum of the scaled squared eigenmappings equals the scaled sum of the squared signal samples:

$$\sum_{k=0}^{N-1} \left(\frac{\hat{z}[k]}{\lambda_k^{1/2}} \right)^2 = \frac{1}{R} \sum_{n=0}^{N-1} |z[n]|^2.$$

Using this result and the fact that $\hat{\mathbf{x}} - \hat{\mathbf{y}}$ is the eigenmapping of the error signal, $x[n] - y[n]$, we obtain:

$$D_{RESIDUAL} \leq \lambda_{N_{VQ}} \left(\frac{R}{N} \sum_{n=0}^{N-1} |x[n] - y[n]|^2 \right).$$

Since the input vector and all output vectors have bounded average power, the final result follows from Equation 4.11 in Theorem 4.2 since the index of the eigenvalue of interest, N_{VQ}, is the the dimension of the vector quantization problem given in Equation 4.14.

Minimizing D_{VQ} in Equation 4.15 given an input eigenmapping $\hat{\mathbf{x}}$ and 2^N output eigenmappings is a classic vector quantization problem [3]. Therefore, for a fixed oversampling ratio, R, there exists a vector quantization problem that is asymptotically equivalent to the oversampled data converter problem as the block size, N, tends to infinity. A new oversampled data converter architecture motivated by this result is presented in the next section.

4.4 The Eigenmodulator

Let the block size, N, the oversampling ratio, R, and the vector quantization dimension, $1 \leq N_{VQ} \leq N$, be fixed. It is convenient to define the $N_{VQ} \times 1$ eigenmapping of a vector $\mathbf{z}$ to be:

$$\hat{\mathbf{z}} = R^{-\frac{1}{2}} \Lambda_{N_{VQ}} \mathbf{V_{N_{VQ}}}^{\dagger} \mathbf{z}$$

where $\Lambda_{N_{VQ}} = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N_{VQ}-1})$ and $\mathbf{V_{N_{VQ}}}$ is formed from the first N_{VQ} columns of $\mathbf{V}$. The eigenvalues λ_k and the unitary matrix $\mathbf{V}$ were defined in Section 4.2. Because of Theorem 4.4, whenever an eigenmapping is referred to in the remainder of this thesis, it denotes an $N_{VQ} \times 1$ vector, unless explicitly stated otherwise.

The set of all legal output vectors is $\mathcal{Q}$. An eigenmodulator is defined to be an oversampled data converter that outputs the vector in $\mathcal{Q}$ whose eigenmapping is closest in a Euclidean sense to the input eigenmapping. This minimizes D_{VQ} in Equation 4.15. Let the vector quantization dimension N_{VQ} be defined as in Equation 4.14 for some $\epsilon > 0$. For a fixed oversampling ratio, R, and a fixed set of legal output vectors, Theorem 4.4 shows that the eigenmodulator is the asymptotically optimal oversampled data converter as the block size, N, tends to infinity.

The eigenmodulator architecture is shown in Fig. 1 and consists of a decimated filter bank and an N_{VQ}-dimensional vector quantizer. The first task of the eigenmodulator is to compute the $N_{VQ} \times 1$ eigenmapping of the input vector. This is accomplished by the N_{VQ}

parallel FIR filters $H_k(z)$, $0 \leq k < N_{VQ}$, where:

$$H_k(z) = R^{-\frac{1}{2}} \lambda_k^{\frac{1}{2}} \sum_{n=0}^{N-1} v_k[N-1-n] z^{-n},$$

and $v_k[n]$ is the (n,k) component of the diagonalizing matrix $\mathbf{V}$. The output of the N-fold decimator [8] on the kth branch is denoted by $\hat{x}_k[n]$. This decimated filter bank breaks the input sequence, $x[n]$, into blocks of length N, and computes the eigenmapping one block at a time. Since this is a multirate system, standard polyphase filtering techniques [8] can be employed to reduce the complexity of the implementation shown in Fig. 1.

The second task of an eigenmodulator is to find and output the vector in $\mathcal{Q}$ whose eigenmapping is closest to the input eigenmapping. The N_{VQ}-dimensional vector quantizer in Fig. 1 can be implemented using N_{VQ} look-up tables, each with $|\mathcal{Q}|$ entries. This is because there are $|\mathcal{Q}|$ legal output vectors and each eigenmapping has N_{VQ} components. In the case where the legal output vectors are binary, $|\mathcal{Q}| = 2^N$, and the memory required by the vector quantizer contains $N_{VQ}2^N$ entries. Because of the exponential dependence of the vector quantizer memory size on the block size, N, it is of interest to find the effects of finite N on the oversampled data converter distortion. At present we are content to bound the distortion using the rate-distortion analysis in the next section.

4.5 Rate-distortion Analysis

This section presents and evaluates a parametric rate-distortion function based on the distortion metric in Lemma 4.1 when the input is a band-limited Gaussian signal. A rate-distortion function [5] for a particular source and fidelity criterion gives the least distortion that can be achieved for a given information rate. The source under consideration is Gaussian. The fidelity criterion is the amount of error power in the desired frequency region after appropriate filtering. For the single-bit data converter, the information rate is unity because exactly one bit is transmitted per input sample. General expressions for arbitrary oversampling ratios and block sizes are obtained and an asymptotic expression valid for large block size is presented.

Consider an input vector, $\mathbf{x}$, whose components are joint Gaussian with zero mean vector and covariance matrix $E\{\mathbf{x}\mathbf{x}^\dagger\} = \Phi = \sigma^2 R\mathbf{A}$. The filter matrix $\mathbf{A}$ is defined in Equation 4.3. Note that the input signal is band-limited over $\Omega = (-\frac{\pi}{R}, \frac{\pi}{R})$ and that its average power is σ^2. For example, if the input process is white then the frequency interval is $\Omega = (-\pi, \pi)$, the oversampling ratio is $R = 1$, the covariance matrix is $\Phi = \sigma^2\mathbf{I}$ and the filter matrix is $\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix.

Consistent with Lemma 4.1, we define our distortion metric, $\rho(\mathbf{x}, \mathbf{y})$, to have the quadratic form:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{N}(\mathbf{y} - \mathbf{x})^\dagger \mathbf{A}(\mathbf{y} - \mathbf{x}).$$

This is to be interpreted as the amount of distortion associated with transmitting output vector $\mathbf{y}$ when the input vector is $\mathbf{x}$. Let $\mathbf{V}$ be the diagonalizing unitary matrix associated with the filter matrix, $\mathbf{A}$, that satisfies

$$\mathbf{V}^\dagger \mathbf{A}\mathbf{V} = \Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \cdots, \lambda_{N-1}),$$

where λ_i is the ith eigenvalue of $\mathbf{A}$, which is associated with the ith eigenvector of $\mathbf{A}$ (which happens to be the ith column of $\mathbf{V}$.) Since $\mathbf{V}$ is unitary, $\mathbf{V}^\dagger\mathbf{V} = \mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$.

Now define the linear transformation

$$\hat{\mathbf{x}} = \Lambda^{1/2}\mathbf{V}^\dagger \mathbf{x},$$

$$\hat{\mathbf{y}} = \Lambda^{1/2}\mathbf{V}^\dagger \mathbf{y}.$$

This transformation is based on the transformation in Section 4.2, without the constant scale factor $R^{-1/2}$. The transformed vector, $\hat{\mathbf{x}}$, is Gaussian because $\mathbf{x}$ is Gaussian and the transformation is linear and non-singular [6]. The transformation is non-singular because none of the eigenvalues of $\mathbf{A}$ are zero [1]. The covariance matrix of the transformed Gaussian vector, $\hat{\mathbf{x}}$, is easily shown to be diagonal:

$$E\{\hat{\mathbf{x}}\hat{\mathbf{x}}^\dagger\} = \Lambda^{1/2}\mathbf{V}^\dagger E\{\mathbf{x}\mathbf{x}^\dagger\}\mathbf{V}\Lambda^{1/2}$$

$$= \sigma^2 R\left(\Lambda^{1/2}\mathbf{V}^\dagger \mathbf{A}\mathbf{V}\Lambda^{1/2}\right) = \sigma^2 R\Lambda^2$$

$$= \text{diag}(\sigma^2 R\lambda_0{}^2, \sigma^2 R\lambda_1{}^2, \cdots, \sigma^2 R\lambda_{k-1}{}^2).$$

Therefore, the components of the transformed vector, $\hat{\mathbf{x}}$, are uncorrelated with respect to each other. And since it is joint Gaussian, each component of $\hat{\mathbf{x}}$ is independent of all other components.

Using this transformation, our distortion metric can now be written in the standard squared-error form:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{N}(\hat{\mathbf{y}} - \hat{\mathbf{x}})^\dagger(\hat{\mathbf{y}} - \hat{\mathbf{x}}).$$

This standard squared-error problem has been solved before [5]. The rate-distortion function can be expressed in parametric form. The information rate and the distortion are expressed in terms of the non-negative parameter θ [5]. The parametric expressions for the information rate and distortion, respectively, are:

$$r(\theta) = \frac{1}{N}\sum_{i=0}^{N-1} \max(0, \frac{1}{2}\log_2(\frac{\sigma^2 R\lambda_i{}^2}{\theta})), \tag{4.16}$$

$$d(\theta) = \frac{1}{N}\sum_{i=0}^{N-1} \min(\theta, \sigma^2 R\lambda_i{}^2). \tag{4.17}$$

A lower case "d" will be used to denote the parametric distortion in order to differentiate it from the distortion metric "D" considered earlier. By the definition of the rate-distortion function, $D \geq d$. By definition, the information rate in the single-bit oversampled data converter is 1 bit/sample. As shown in Section 4.2, the eigenvalues are non-zero. Therefore, the linear transformation that generates $\hat{\mathbf{x}}$ from $\mathbf{x}$ is invertible and the information rate is preserved [5]. Therefore, the present task is to find the value of θ_1 such that $r(\theta_1) = 1$. This value exists and is unique [5]. The distortion corresponding to $d(\theta_1)$ is therefore the minimum distortion that can be achieved by a single-bit oversampled data converter. A simple gradient search method was employed to find the parameter value θ_1 such that $r(\theta_1) = 1$.

Figures 2 and 3 present the single-bit distortion bounds as a function of the oversampling ratio and the block size. The block size is a measure of system complexity. The results of this section should be viewed as a reference for future efforts involving vector quantizers in oversampled data converters. As all realizable systems must have finite complexity, the

effects of finite block size are important. As intuitively expected, the distortion decreases as the oversampling ratio increases. For a fixed oversampling ratio, the distortion bound tends to a limit as the block size becomes large. This will be discussed shortly. It is interesting to note that the distortion does not necessarily decrease monotonically with the block size when the block size is small. There is a ripple in each of the curves. This is because as the block size increases, there will be more eigenvalues tending to unity. The ripple is inherent in the rate-distortion function and is not a computational artifact. The mechanism for the ripple is best seen by considering the evaluation of the parametric rate-distortion function. Given the block size and the oversampling ratio, the eigenvalues of $\mathbf{A}$ are first solved for. Then the parameter θ_1 is obtained that satisfies $r(\theta_1) = 1$ in Equation 4.16. When the number of eigenvalues that are close to unity increases by one, θ_1 must become larger in order to preserve $r(\theta_1) = 1$. This causes an increase in the value of $d(\theta_1)$ as seen in Equation 4.17. The correlation between the ripple and the number of non-zero terms in Equation 4.16 is shown in Fig. 4. Terms in Equation 4.16 are non-zero when $\sigma^2 R\lambda_i$ exceed θ. The eigenvalues for which this is true are referred to as "large" in Fig. 4. Consistent with this analysis, as the block size becomes large, the ripple becomes less noticeable, since the presence of one more eigenvalue that is close to unity has a diminishing effect on r and d when there are already many eigenvalues close to unity.

As the block size becomes large, the minimum distortion asymptotically tends to a constant. This will be shown in two ways. The first way involves the Toeplitz distribution theorem and gives a nice spectral interpretation. The second way does not require knowledge of the Toeplitz distribution theorem and instead uses only the available knowledge about the asymptotic behavior of the eigenvalues. Of course, the two approaches are not independent.

The first approach uses the Toeplitz distribution theorem [7] and the corresponding integrals that follow from the parametric rate-distortion expressions as detailed by Berger [5]. These integrals make use of the spectrum of the stationary input process:

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \phi_k e^{-j\omega k},$$

where ϕ_k is the entry on the kth diagonal of the infinite Toeplitz matrix, Φ_∞, which is the autocorrelation function with lag k. In our band-limited Gaussian example, the spectrum is:

$$\Phi(\omega) = \begin{cases} \sigma^2 R, & \omega \in \Omega = (-\frac{\pi}{R}, \frac{\pi}{R}) \\ 0, & \text{otherwise.} \end{cases}$$

As a consequence of the Toeplitz distribution theorem, the parametric information rate and distortion expressions can be written as:

$$r(\theta) = \frac{1}{2}\int_{-\pi}^{\pi} \max[0, \log(\frac{\Phi(\omega)}{\theta})]\frac{d\omega}{2\pi},$$

$$d(\theta) = \int_{-\pi}^{\pi} \min[\theta, \Phi(\omega)]\frac{d\omega}{2\pi},$$

where the parameter θ is non-negative [5]. It easily follows that:

$$r(\theta) = \frac{1}{2}\int_{\Omega} \max[0, \log(\frac{\sigma^2 R}{\theta})]\frac{d\omega}{2\pi} = \frac{1}{2R}\max[0, \log(\frac{\sigma^2 R}{\theta})]$$

since the above integrand is independent of frequency. It is also true that:

$$d(\theta) = \int_{\Omega} \min[\theta, \sigma^2 R]\frac{d\omega}{2\pi} = \begin{cases} \frac{\theta}{R}, & 0 \le \theta < \sigma^2 R \\ \sigma^2, & \text{otherwise.} \end{cases}$$

There are two cases. When $\theta \ge \sigma^2 R$, $d = \sigma^2$ and $r = 0$. This is when the permissible distortion is as large or larger than the signal power, in which case, no information needs to be sent. The other case is when $0 \le \theta < \sigma^2 R$. Then $d = \theta/R < \sigma^2$ and we can express the information rate as a function of the permissible distortion: $r(d) = \frac{1}{2R}\log(\frac{\sigma^2}{d})$. Setting the information rate to unity and rearranging, we obtain the final result for the minimum distortion as a function of the oversampling ratio as the block size tends to infinity when the input is a band-limited Gaussian:

$$d = \sigma^2 2^{-2R}, \tag{4.18}$$

where σ^2 is the power of the input signal and R is the oversampling ratio.

This asymptotic behavior can also be shown using our knowledge of the eigenvalues and their asymptotic behavior. The parametric expression for the distortion in Equation 4.17 may be written as:

$$d(\theta) = \frac{N_1}{N}\theta + \frac{1}{N}\sum_{k=N_1}^{N-1} R\sigma^2\lambda_k^2, \tag{4.19}$$

where N_1 is the number of eigenvalues such that $\sigma^2 R$ times the square of the eigenvalue exceeds θ. We may similarly write the parametric expression for the information rate as:

$$r(\theta) = \frac{1}{N} \sum_{k=0}^{N_1-1} \frac{1}{2} \log(\frac{\sigma^2 R \lambda_k^2}{\theta}).$$

As the block size, N, tends to infinity, the eigenvalues will either tend to unity or zero in a manner described by Theorem 4.2. Therefore, the eigenvalues in the sum over k in Equation 4.19 must tend to zero. This results in Equation 4.18 also.

The in-band MSE distortion as a function of oversampling ratio is perhaps the primary metric of oversampled data converters. The distortion in conventional $\Delta\Sigma$ converters is polynomial with respect to the oversampling ratio, while the rate-distortion bound is exponential with respect to the oversampling ratio. The disparity between the rate-distortion bound and the $\Delta\Sigma$ performance is not surprising in light of the fact that the $\Delta\Sigma$ modulator is a relatively simple device. Rate-distortion theory is useful in obtaining distortion bounds but it does not generally indicate how tight these bounds will be for particular systems.

Appendix I: Asymptotic Evaluation of Slepian Constants

This appendix asymptotically evaluates Equations 43 and 47 in Slepian's paper [1] and applies them to Equations 59 and 63 in [1] in order to prove Theorem 4.3. The quantity ϵ is assumed to be fixed. The quantities N, R and N/R are assumed to tend to infinity. The condition that N and N/R tend to infinity is necessary in order to use Equation 59 and 63 in [1]. Note that Slepian's bandwidth constant, W, is related to our oversampling ratio, R, by $W = \frac{1}{2R}$.

Case 1: The eigenvalue with index $k = \lfloor \frac{N}{R}(1-\epsilon) \rfloor$

Without loss of generality, let ϵ be small. If it is not small, then we can find a small ϵ' to make a new $k' = \lfloor \frac{N}{R}(1-\epsilon') \rfloor$. Then we can bound $1 - \lambda_k$ using the fact that the eigenvalues

are decreasing in order so that $1-\lambda_k < 1-\lambda_{k'}$, since $k' > k$.

From Equation 59 of [1],

$$1-\lambda_k \sim e^{-CL_4/2}e^{-L_3N} \leq e^{-L_3N},$$

where the inequality comes from the fact that $CL_4/2$ is a positive quantity based on Equations 45 and 47 in [1]. Since L_3 depends on B, we must first solve for the value of B that satisfies its defining integral in Equation 43 of [1]:

$$\int_B^1 \sqrt{\frac{\xi - B}{(\xi - A)(1-\xi^2)}}d\xi = \frac{k\pi}{N},$$

where $A = \cos(\frac{\pi}{R})$. We begin with the estimate $B = \cos(\frac{\pi(1-\mu)}{R})$, where μ is assumed to be small. We then compute μ and verify our assumption. Changing variables so that $\xi = \cos(\frac{\pi\theta}{R})$, our integral expression becomes

$$\frac{\pi}{R}\int_0^{1-\mu} \sqrt{\frac{\cos(\frac{\pi\theta}{R}) - \cos(\frac{\pi(1-\mu)}{R})}{\cos(\frac{\pi\theta}{R}) - \cos(\frac{\pi}{R})}}d\xi, = \frac{k\pi}{N} \sim \frac{\pi(1-\epsilon)}{R}.$$

The integrand is well-defined over the interval $[0, 1-\mu]$. In the limit of large R, we may express the integrand using the expansion $\cos(x) = 1 - \frac{x^2}{2} + O(x^4)$. Dropping the order terms, the integrand is asymptotically equal to $((1-\mu)^2 - \theta^2)^{1/2}(1-\theta^2)^{-1/2}$. Making a final change of variables so that $u = \frac{\theta}{1-\mu}$, our integral expression can asymptotically be expressed as:

$$\frac{\pi(1-\mu)}{R}\int_0^1 \left(\frac{1-u^2}{\frac{1}{(1-\mu)^2} - u^2}\right)^{1/2} du \sim \frac{\pi(1-\epsilon)}{R}.$$

As μ becomes small, the integrand tends to unity for all points on the open interval $[0, 1)$. Therefore, in the limit of small μ, the above integral tends to unity and

$$\frac{\pi(1-\mu)}{R} \sim \frac{\pi(1-\epsilon)}{R}.$$

Therefore $\mu \sim \epsilon$, which verifies our earlier assumption since ϵ is small. Therefore, in the limit of large R and small ϵ, $B \sim \cos(\frac{\pi(1-\epsilon)}{R}) \sim \cos(\frac{\pi k}{N})$.

The quantity L_3 is defined in Equation 47 of [1]:

$$L_3 = \int_A^B \left|\frac{\xi - B}{(\xi - A)(1-\xi^2)}\right|^{1/2} d\xi.$$

Change variables as before so that $\xi = \cos(\frac{\pi\theta}{R})$. The integrand, $I(\theta)$, is singular at $\theta = 1$. Rather than deal with the singularity (it can be shown that the integral is still convergent), we note that the integrand is positive so that:

$$L_3 \geq L_3^* \equiv \int_{1-\epsilon}^{1-\frac{\epsilon}{2}} I(\theta)d\theta.$$

Therefore, $e^{-L_3 N} \leq e^{-L_3^* N}$, which can be used to generate an asymptotic bound on $1 - \lambda_k$.

The integrand is well-defined over the interval $[1-\epsilon, 1-\frac{\epsilon}{2}]$. Again for large R, we expand the cosine terms in the integrand and ultimately show that $I(\theta) \sim |(1-\epsilon)^2 - \theta^2|^{1/2}|1-\theta^2|^{-1/2}$, as before. Make the final variable change, $u = \frac{2}{\epsilon}(\theta - 1 + \epsilon)$ so that $1 - \epsilon$ maps to zero and $1 - \epsilon/2$ maps to unity. Then in the limit of small ϵ our integrand tends to $u^{1/2}(2-u)^{-1/2}$ on the open interval $(0, 1]$ (at $u = 0$, the integrand has a first-order zero). We can lower bound the integrand by $\frac{u^{1/2}}{2}$ over the interval $0 < u \leq 1$. Therefore,

$$L_3^* \geq \frac{\epsilon\pi}{2R}\int_0^1 \frac{u^{1/2}}{2}du = \frac{\epsilon\pi}{6R}.$$

This gives the final result that $1 - \lambda_k \leq \exp(-\frac{\epsilon\pi N}{6R})$ and the desired order expression.

Case 2: The eigenvalue with index $k = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$

Now we consider the behavior of λ_k when the index is $k = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$. Let ϵ be fixed (but not necessarily small as in the analysis for $k = \lfloor \frac{N}{R}(1-\epsilon) \rfloor$ in the previous subsection) and let $R \to \infty$.

As discussed by Slepian [1], we use the eigenvalue symmetry about the index N/R. To find information on $\lambda_k(N, R)$, we study $\lambda_{k'}(N, R')$ where $k' = N - k - 1$ and $\frac{1}{R'} = 1 - \frac{1}{R}$. Then $\lambda_k(N, R) = 1 - \lambda_{k'}(N, R')$ (Equation 13 in [1]). We can write $k' = \lfloor \frac{N}{R'}(1-\epsilon') \rfloor$, where $\epsilon' \sim \frac{\epsilon}{R}$. Since $R \to \infty$, $R' \to 1$ and $\epsilon' \to 0$, so we cannot use the asymptotic analysis for large oversampling ratios in the preceding subsection. From Equation 59 in [1],

$$1 - \lambda_{k'} \sim e^{-CL_4/2} e^{-L_3 N} \leq e^{-L_3 N},$$

where the inequality follows from the fact that CL_4 is strictly non-negative (see Equation 45 and Equation 47 in [1]). In order to compute L_3, we need the quantity B that satisfies

Equation 43 in [1]:

$$\int_B^1 \sqrt{\frac{\xi - B}{(\xi - A)(1-\xi^2)}} d\xi = \frac{k'\pi}{N} \sim \frac{\pi(1-\epsilon')}{R'} \sim \pi(1 - \frac{\epsilon}{R}),$$

where $A = \cos(\frac{\pi}{R'}) \sim \cos(\pi(1 - \frac{1}{R}))$. We make the initial estimate that $B \sim \cos(\pi(1-\mu))$ and assume μ is small. We then solve for μ to verify our assumption. Make the change of variables $\xi = \cos(\frac{\pi(1-\theta)}{R'})$ and then set $u = \frac{\theta - \mu}{1-\mu}$. The integral equation for B becomes

$$\frac{\pi(1-\mu)}{R'} \int_0^1 I(u) du \sim \frac{\pi(1-\epsilon')}{R'}.$$

For $R \to \infty$, the above integrand converges to unity at all points on the open interval $(0, 1]$. Therefore, $B \sim \cos(\frac{k'\pi}{R'})$ and $\mu \sim \epsilon' \sim \frac{\epsilon}{R} \to 0$, verifying our assumption that μ is small.

Now we evaluate L_3 in Equation 47 of [1]:

$$L_3 = \int_A^B \left| \frac{\xi - B}{(\xi - A)(1-\xi^2)} \right|^{1/2} d\xi.$$

Change variables so that $\xi = \cos(\frac{\pi(1-\theta)}{R'})$. The resulting integrand, $I(\theta)$, is singular at $\theta = 0$. We note that the integrand is non-negative. As before, we now consider a L_3^* such that:

$$L_3 \geq L_3^* \equiv \frac{\pi}{R'} \int_{\epsilon'/2}^{\epsilon'} I(\theta) d\theta,$$

and avoid the singularity. We still preserve the useful bound $e^{-L_3 N} \leq e^{-L_3^* N}$. Change variables to make $u = \frac{2}{\epsilon'}(\theta - \epsilon'/2)$. Then we have:

$$L_3^* = \frac{\pi\epsilon'}{2R'} \int_0^1 f(u) du,$$

where the new integrand, $f(u)$, converges to unity in the open interval $(0, 1]$ as R goes to infinity. Finally, we have the result that

$$L_3^* \sim \frac{\pi\epsilon'}{2R'} \sim \frac{\pi\epsilon}{2R}.$$

This is combined with the eigenvalue symmetry discussed earlier and the bound $1 - \lambda_{k'} \leq e^{-L_3^* N}$ to obtain the desired inequality

$$\lambda_k = O(\exp(-\frac{\alpha N}{R})),$$

for $\alpha = \frac{\pi\epsilon}{2}$ and $k = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$.

Appendix II: An Eigenvalue Upper Bound

Lemma A.2. *The eigenvalues of the filter matrix,* $\mathbf{A}$*, in Equation 4.3 satisfy the inequality*

$$\lambda_k \leq 1, \tag{A.2.1}$$

for all $k = 0, 1, \ldots, N-1$ *and all* $R \geq 1$*. Equality occurs if and only if the oversampling ratio,* R*, is unity.*

Proof: The characteristic equation for the eigenvectors, $\mathbf{v}_k$, and eigenvalues, λ_k, of the $N \times N$ filter matrix $\mathbf{A}$ is:

$$\lambda_k \mathbf{v}_k = \mathbf{A}\mathbf{v}_k, \tag{A.2.2}$$

where $k = 0, 1, \ldots, N-1$. The eigenvalues and eigenvectors are ordered so that $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{N-1}$. Using the definition of the filter matrix in Equation 4.3, the characteristic equation may be written as a frequency domain integral:

$$\lambda_k v_k[n] = \int_{-\pi/R}^{\pi/R} V_k(e^{jw}) e^{jwn} \frac{d\omega}{2\pi}, \tag{A.2.3}$$

where $v_k[n]$ is the nth component ($n = 0, 1, ..., N-1$) of the kth eigenvector, $\mathbf{v}_k$, and $V_k(e^{jw})$ is defined as its Fourier transform:

$$V_k(e^{jw}) = \sum_{n=0}^{N-1} v_k[n] e^{-jwn}. \tag{A.2.4}$$

Multiply both sides of Equation A.2.3 by $v_k[n]$ and sum from $n = 0$ to $n = N-1$. Using the fact that the eigenvectors are orthonormal, we obtain:

$$\lambda_k = \int_{-\pi/R}^{\pi/R} V_k(e^{jw}) \left(\sum_{n=0}^{N-1} v_k[n] e^{jwn} \right) \frac{d\omega}{2\pi} = \int_{-\pi/R}^{\pi/R} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi}. \tag{A.2.5}$$

Combining Parseval's relation with the orthonormality of the eigenvectors leads to the expression that:

$$\int_{-\pi}^{\pi} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi} = \sum_{n=0}^{N-1} v_k^2[n] = 1. \tag{A.2.6}$$

Finally, we combine Equations A.2.5 and A.2.6 to conclude our proof:

$$\lambda_k = \int_{-\pi/R}^{\pi/R} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi} \leq \int_{-\pi}^{\pi} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi} = 1, \tag{A.2.7}$$

since equality follows if and only if $R = 1$. As a check, when $R = 1$, Equation 4.3 shows that the filter matrix $\mathbf{A}$ is the $N \times N$ identity matrix, which has all N eigenvalues equal to unity. ∎

References

1. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty - V: The Discrete Case," *Bell Systems Technical Journal,* vol. 57, pp. 1371-1430, May-June 1978.

2. J. Huang, P. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables," *IEEE Transactions on Communications Systems*, vol. CS-11, pp. 289-296, September 1963.

3. A. Gersho and R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, 1991.

4. J. Franklin, Matrix Theory, Prentice-Hall, Englewood Cliffs, NJ, 1968.

5. T. Berger, Rate Distortion Theory: A Mathematical Basis for Data Compression, Prentice-Hall, Englewood Cliffs, NJ, 1971.

6. W. Feller, An Introduction to Probability Theory and Its Applications, Volume II, Wiley, New York, 1966.

7. U. Grenander, G. Szegö, Toeplitz Forms and their Applications, Univ. of California Press, Berkeley and Los Angeles, CA, 1958.

8. P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice-Hall, New York, 1993.

Figures

Fig. 1: The Eigenmodulator

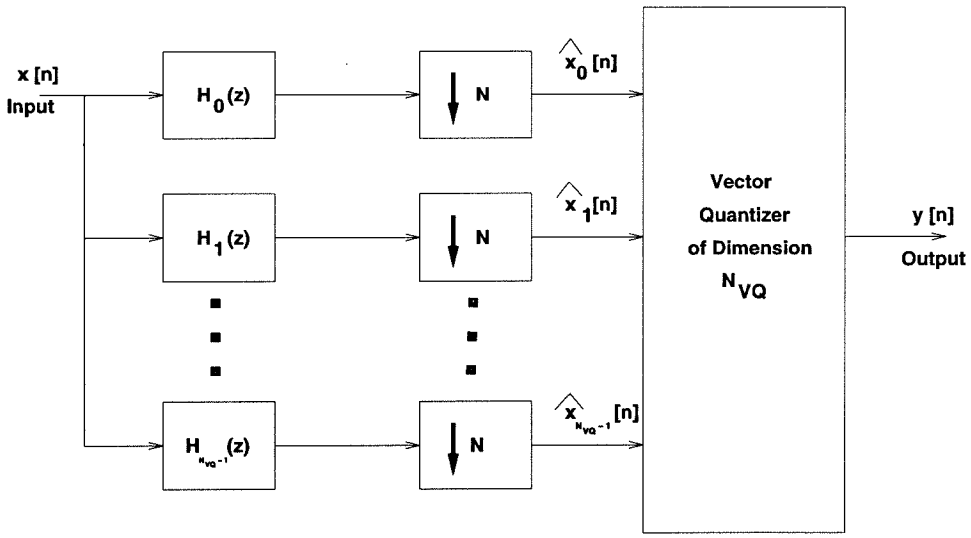


Fig. 2: Single-bit Distortion Bounds versus Block size and Oversampling ratio

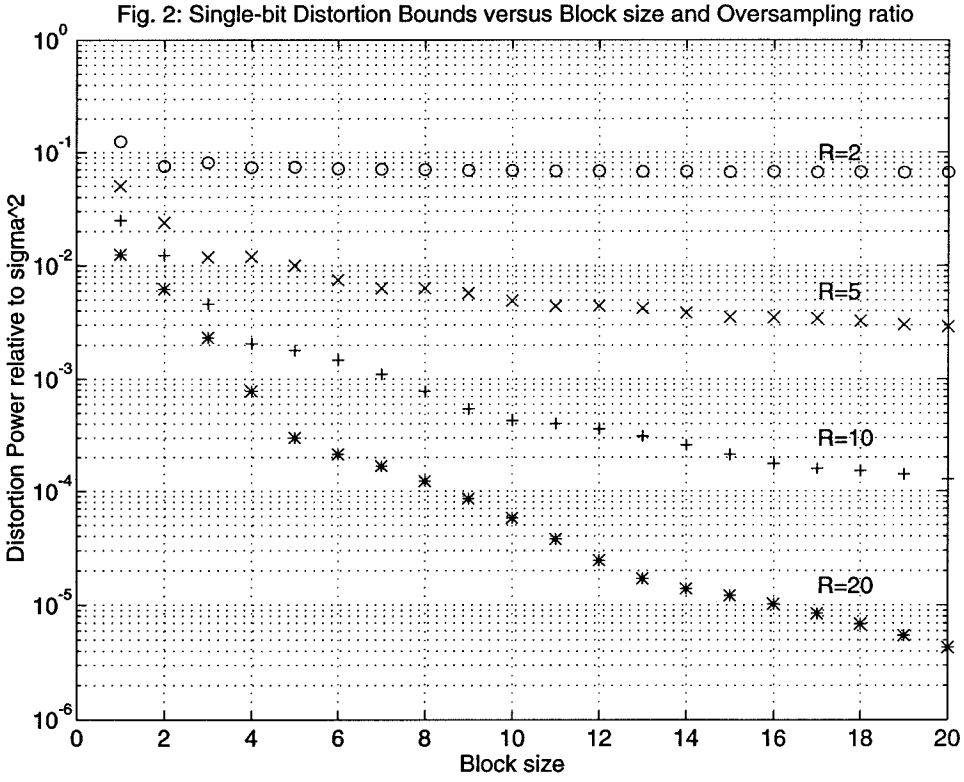
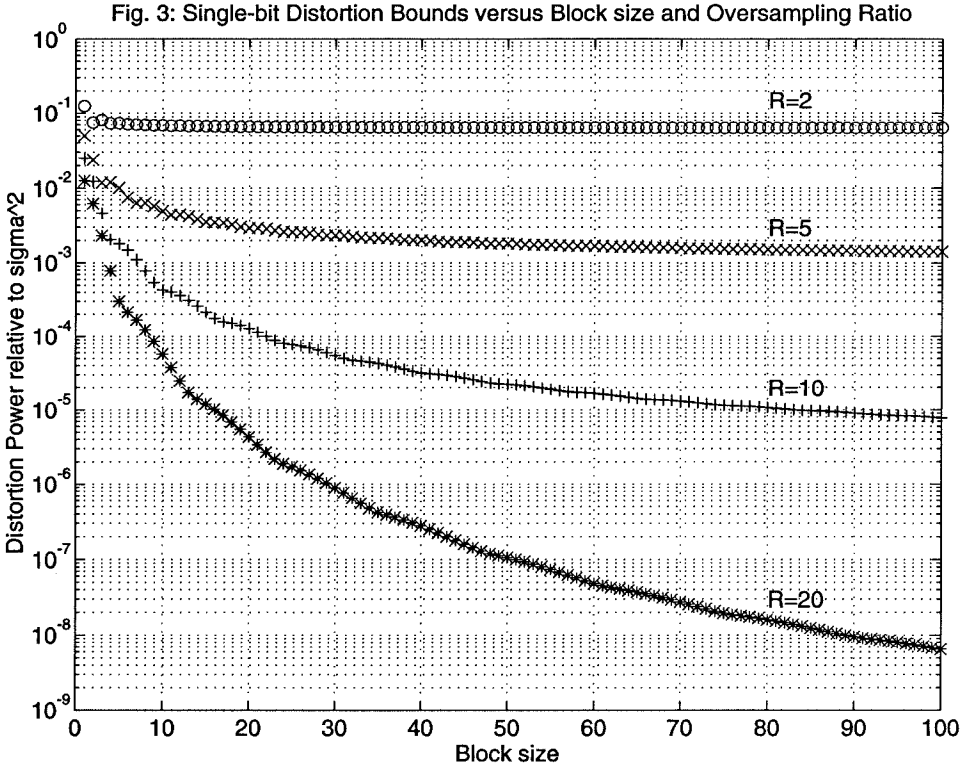


Fig. 3: Single-bit Distortion Bounds versus Block size and Oversampling Ratio



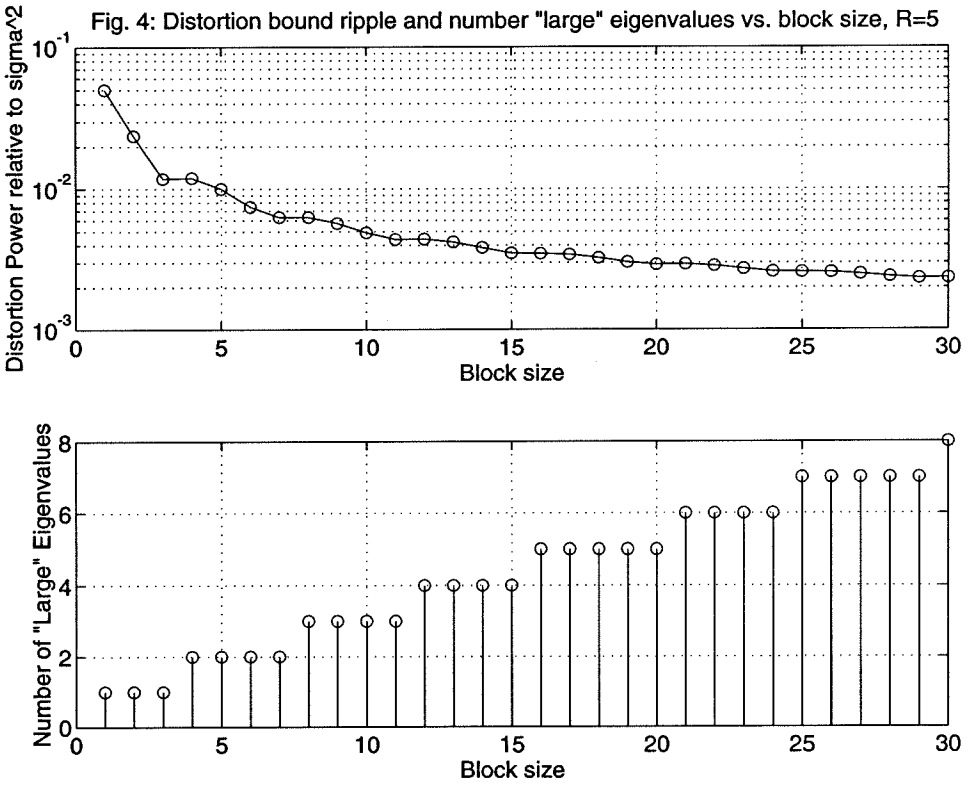


Fig. 4: Distortion bound ripple and number "large" eigenvalues vs. block size, R=5
Distortion Power relative to sigma^2
10^-1
10^-2
10^-3
0
5
10
15
20
25
30
Block size
Number of "Large" Eigenvalues
0
2
4
6
8
0
5
10
15
20
25
30
Block size

Chapter 5

The Single-Bit Eigenmodulator in an Asymptotic Scenario

5.1 Introduction and Results

The objective of this chapter is to study the nature of the output eigenmappings in a framework suitable for application of vector quantization theory. A centroid condition is one of the three known necessary conditions for an optimal vector quantizer [1] and is detailed in Section 5.6. Under the conditions of Theorem 4.4, it follows that the eigenmodulator with the optimal vector quantizer will be the optimal oversampled data converter. The centroid condition requires information about the distribution of the output eigenmappings, which we pursue in an asymptotic scenario.

The important case of the single-bit data converter is analyzed exclusively. The complexity of the eigenmodulator in the previous chapter is driven by the complexity of its internal vector quantizer. It was shown that when the vector quantizer in the single-bit eigenmodulator employs an exhaustive look-up table, there are $N_{VQ}2^N$ memory entries. N_{VQ} denotes the dimension of the vector quantizer, where $1 \leq N_{VQ} \leq N$ and N is the block size. The previous chapter investigated the effects of fixed and asymptotically large block sizes through a rate-distortion analysis without giving information about the nature of the eigenmappings.

This chapter studies the eigenmappings when the dimension of the vector quantizer is small compared to the oversampling ratio, R, in an asymptotic scenario where, in addition to the block size, the oversampling ratio becomes large. It is shown that the distribution of the output eigenmappings is asymptotically joint Gaussian when $\frac{N_{VQ}^2}{R}$ tends to zero. This fact combined with the necessary centroid condition for an optimal vector quantizer [1] motivates the study of the eigenmodulator when the input signal is stationary, band-limited and joint Gaussian. Vector quantization theory is used to comment on the output scale factor for Gaussian input signals in this asymptotic scenario.

We can ensure that $\frac{N_{VQ}^2}{R}$ tends to zero by making N_{VQ} a function of the oversampling ratio R, e.g., $N_{VQ} = R^{\frac{1}{2}-\eta}$ for $0 < \eta < \frac{1}{2}$. This can have implications for the block size, N. For large $N_{VQ} = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$, with $\epsilon > 0$, the residual distortion is $O(\exp(-\frac{\alpha N}{R}))$ as shown in Theorem 4.3, where α is a positive constant. In this case, the rate at which the block size, N, tends to infinity is a function of the oversampling ratio, R, in order to ensure that the distribution of the output eigenmappings is asymptotically joint Gaussian. In the above example, the block size is approximately $R^{\frac{3}{2}-\eta}$. In contrast, for a fixed oversampling ratio, the rate-distortion analysis of the previous chapter and our intuition suggests that the average distortion decreases as we make the block size larger in a manner independent of the oversampling ratio. However, the distribution of the eigenmappings would then not necessarily be joint Gaussian; the exact distribution of the eigenmappings when the block size tends to infinity at a rate uncoupled from the oversampling ratio is an open research problem. Since knowledge of the distribution of output vectors can be used to facilitate the design of efficient vector quantizers [1], constraining the rate at which the block size grows to ensure the joint Gaussian result is of practical, as well as theoretical, interest. The results of this chapter lays the groundwork for future analysis and design of oversampled data converters using vector quantization theory.

The next section equates the density of the output vector eigenmappings with the density of a suitably defined random vector process. Using this result, the marginal and joint densities of the binary ($\pm c$) output eigenmappings are shown to become Gaussian under certain

asymptotic conditions. A generalized Kolmogorov-Smirnov test [2] is used to examine the joint density result using realistic values for the oversampling ratio and vector quantization dimension. This knowledge of the joint density is finally combined with vector quantization theory to comment on the asymptotic output scaling of the eigenmodulator when the input is a band-limited Gaussian process.

5.2 The Density of the Binary Output Vector Eigenmappings

This section shows that the distribution of the eigenmappings of all binary output vectors in N_{VQ}-dimensional space can be related to the probability density function (p.d.f.) of a suitably defined random vector. For a given output vector, $\mathbf{y}$, the $N_{VQ} \times 1$ eigenmapping, $\hat{\mathbf{y}}$, is defined as:

$$\hat{\mathbf{y}} = R^{-\frac{1}{2}} \Lambda_{N_{VQ}}^{1/2} \mathbf{V}_{N_{VQ}}^{\dagger} \mathbf{y}, \tag{5.1}$$

where $\Lambda_{N_{VQ}}$ and $\mathbf{V}_{N_{VQ}}$ were defined in the previous section.

There are 2^N binary output vectors of length N. Each component of every output vector is either $+c$ or $-c$. This constant with respect to time, c, is called the *output scale value* and is important in systems where the data converter output is fed back to or compared with the converter input. Any constant offset common to both binary output levels is ignored at present because the output is linear and, without loss of generality, it is assumed that the input has zero mean. If instead the stationary input process has some non-zero mean we may subtract this constant from the input prior to the data converter and add the constant at the output of the data converter.

Define a random vector, $\mathbf{z} = [z[0]z[1]\ldots z[N-1]]^{\dagger}$, to have components that are independent and identically distributed (i.i.d.) such that:

$$z[n] = \begin{cases} +c & \text{with probability } 1/2, \\ -c & \text{with probability } 1/2, \end{cases} \tag{5.2}$$

for $n = 0, 1, ..., N-1$. The random vector $\mathbf{z}$ assumes the value of each output vector with

equal probability, 2^{-N}. We can define the eigenmapping of the random vector in a similar fashion to that above:

$$\hat{\mathbf{z}} = R^{-1/2}\Lambda_{N_{VQ}}^{1/2}\mathbf{V}_{N_{VQ}}^{\dagger}\mathbf{z}. \tag{5.3}$$

This will be referred to as the *random eigenmapping.* The components of the random eigenmapping, $\hat{\mathbf{z}}$, have a well-defined joint cumulative distribution function. Since the random vector is discretely distributed, the random eigenmapping is discretely distributed as well, and the joint probability density function (joint p.d.f.) is defined only when we use Dirac delta functions. Noting this technicality, we permit the use of the joint p.d.f.

By definition of the p.d.f., the probability that a random eigenmapping is located in some region Γ of N_{VQ}-dimensional space, is

$$\text{Prob } \{ \hat{\mathbf{z}} \in \Gamma \} = \int_{\Gamma} q(\hat{\mathbf{z}}) d\hat{\mathbf{z}}, \tag{5.4}$$

where $q(\hat{\mathbf{z}})$ is the N_{VQ}-dimensional joint p.d.f. of the random eigenmapping, $\hat{\mathbf{z}}$. Since each of the 2^N equally-likely random vectors creates an eigenmapping, the probability that the random eigenmapping lies in Γ equals the number of realizations of $\mathbf{z}$ that make an eigenmapping that lies in Γ times the probability of each realization, 2^{-N}. But the number of realizations of $\mathbf{z}$ that make an eigenmapping that lies in Γ equals the number of output vectors that make an eigenmapping that lies in Γ, which we define to be ν_{Γ}. This gives the desired result equating the number of eigenmappings of the output vectors in some region of space Γ to the joint p.d.f. of the random eigenmapping integrated over Γ times 2^N:

$$\nu_{\Gamma} = 2^N \int_{\Gamma} q(\hat{\mathbf{z}}) d\hat{\mathbf{z}}. \tag{5.5}$$

At this time it is convenient to introduce vector quantization terminology. For each output vector, $\mathbf{y}_k$, $k = 0, 1, ..., 2^N - 1$, the *nearest-neighbor region*, S_k, is defined to be that region of N_{VQ}-dimensional space closer in the Euclidean sense to the eigenmapping of $\mathbf{y}_k$ than to the eigenmapping of any other output vector:

$$S_k = \{\hat{\mathbf{x}} \in \mathcal{R}^{N_{VQ}} : d(\hat{\mathbf{x}}, \hat{\mathbf{y}}_k) \leq d(\hat{\mathbf{x}}, \hat{\mathbf{y}}_m), \quad \forall m \neq k\}, \tag{5.6}$$

where the mean-Euclidean distance operator is defined as:

$$d(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \frac{1}{N_{VQ}} \sum_{k=0}^{N_{VQ}-1} (\hat{x}[k] - \hat{y}[k])^2.$$

Combining the asymptotically equivalent vector quantization problem in Theorem 4.3 and the above definition yields an important, but straightforward, result with regards to minimizing distortion metric described in Equation 4.1 of the previous chapter. If the eigenmapping of the input vector lies in the kth nearest-neighbor region, then the kth output vector should be selected. The implication of this result for the design of VQ-based single-bit data converters was considered in the previous chapter.

Using the same random vector approach, it is easy to show that the locations of the eigenmappings of the output vectors satisfy a useful centroid condition. Centroids figure prominently in modern vector quantization analysis [1,4,5]. The *centroid* of the kth nearest-neighborhood, S_k, with respect to some joint p.d.f. is defined to be the expected value of the random vector described by the joint p.d.f. given that the random vector is an element of S_k. This definition leads to the following theorem.

Theorem 5.1. *The eigenmapping of the kth output vector is the centroid of the kth nearest-neighborhood, S_k given the joint p.d.f. $q(\hat{\mathbf{z}})$:*

$$\hat{\mathbf{y}}_k = \frac{\int_{S_k} \hat{\mathbf{z}}\, q(\hat{\mathbf{z}}) d\hat{\mathbf{z}}}{\int_{S_k} q(\hat{\mathbf{z}}) d\hat{\mathbf{z}}}.$$

Proof: The joint p.d.f. is $q(\hat{\mathbf{z}})$, defined in Equation 5.4. There is a one-to-one correspondence between the probability space of the random eigenmapping described in Equation 5.3 and the set of output eigenmappings. For the random eigenmapping described in Equation 5.3, there is exactly one unique realization that resides in S_k, namely $\hat{\mathbf{y}}_k$. This is because if there were two or more distinct realizations in S_k, then points in an arbitrarily small neighborhood of $\hat{\mathbf{y}}_l$, where $l \neq k$, would be closer to $\hat{\mathbf{y}}_l$ than $\hat{\mathbf{y}}_k$, violating the definition of the kth nearest neighborhood region in Equation 5.6. Therefore, the expected value of $\hat{\mathbf{z}}$ given that $\hat{\mathbf{z}} \in S_k$ is $\hat{\mathbf{y}}_k$. ∎

This centroid condition will prove useful in a later section that applies vector quantization theory to study the optimal output scale value, c, for the vector quantization problem

associated with the oversampled data converter problem. The centroid condition is also of general interest because it gives information about the geometric nature of the vector quantizer [1].

5.3 The Marginal P.D.F.

Theorem 5.2. *As the oversampling ratio, R, becomes large, the marginal p.d.f. of the kth component of the binary ($\pm c$) output eigenmapping, $k = 0, 1, \ldots, N-1$, tends to a Gaussian p.d.f. with zero mean and variance $\frac{c^2}{R}\lambda_k$.*

Comment: Theorem 11 in Cramér's classic text [6] states that a necessary and sufficient condition for the convergence of a sequence of p.d.f.'s to a limiting p.d.f. is that the corresponding sequence of characteristic functions converge to the limiting characteristic function uniformly in every finite interval. We therefore consider the characteristic function of the binary output eigenmappings as a function of the oversampling ratio, R, in light of the results of Section 5.2.

Proof: Based on Equation 5.3, the kth component of the random eigenmapping $\hat{\mathbf{z}}$ can be written as:

$$\hat{z}[k] = R^{-1/2}\lambda_k^{1/2} \sum_{n=0}^{N-1} v_k[n]z[n], \tag{5.7}$$

where $k = 0, 1, \ldots, N-1$, and the i.i.d. $z[n]$ sequence is defined in Equation 5.2. The marginal characteristic function of $\hat{z}[k]$ is defined as the Fourier transform of its marginal p.d.f.:

$$\Phi_{\hat{z}[k]}(\xi) = E\{\exp(j\xi\hat{z}[k])\} = \prod_{n=0}^{N-1} E\{\exp(j\xi R^{-\frac{1}{2}}\lambda_k^{1/2}v_k[n]z[n])\},$$

where the last equality follows from the that fact that the $z[n]$ sequence is i.i.d. From the probability mass function of $z[n]$ in Equation 5.2 it follows that $E\{\exp(juz[n])\} = \cos(cu)$. Therefore, the marginal characteristic function of $\hat{z}[k]$ is:

$$\Phi_{\hat{z}[k]}(\xi) = \prod_{n=0}^{N-1} \cos(cR^{-\frac{1}{2}}\lambda_k^{1/2}v_k[n]\xi).$$

Let $cR^{-\frac{1}{2}} = \gamma$, which is assumed to be fixed with respect to the oversampling ratio, R, and the block size, N. A later section addressing the scale factor, c, will justify this assumption. Using the result of Appendix I, the argument of each cosine in the above product can be bounded:

$$|\gamma\xi\lambda_k^{1/2}v_k[n]| \leq \frac{|\gamma||\xi|\lambda_k^{1/2}}{\lambda_k^{1/2}R^{\frac{1}{2}}} \leq \frac{|\gamma||\xi|}{R^{\frac{1}{2}}}.$$

For fixed γ and $|\xi| \leq a$ where a is an arbitrary, fixed constant, the argument goes to zero as $R \rightarrow \infty$. We may then take the natural logarithm of the characteristic function since it will be well-defined as $R \rightarrow \infty$:

$$\ln(\Phi_{\hat{z}[k]}(\xi)) = \sum_{n=0}^{N-1} f(\gamma\lambda_k^{1/2}v_k[n]\xi), \tag{5.8}$$

where the function $f(x) = \ln(\cos(x))$. The Taylor Series expansion about the origin of $f(x)$ with a fourth-order remainder can be written as:

$$f(x) = -\frac{1}{2}x^2 + r_4(x),$$

where the remainder term is:

$$r_4(x) = \frac{f^{(4)}(\zeta)x^4}{4!},$$

and ζ is a point that lies between zero and x. The fourth derivative of $f(x)$ is

$$f^{(4)}(\zeta) = \frac{-2 - 6\sin^2(\zeta)}{\cos^5(\zeta)},$$

which can be bounded by some constant β as the oversampling ratio tends to infinity. This is because we have shown that the argument of each function, f, in Equation 5.8 tends to zero and thus ζ tends to zero. When ζ tends to zero, $f^{(4)}(\zeta)$ can be bounded by a constant. Using this expansion, we can write Equation 5.8 as:

$$\ln(\Phi_{\hat{z}[k]}(\xi)) = -\frac{1}{2}\gamma^2\xi^2\lambda_k \sum_{n=0}^{N-1} v_k^2[n] + \sum_{n=0}^{N-1} r_4(\gamma\lambda_k^{1/2}v_k[n]\xi). \tag{5.9}$$

The eigenvectors were shown to be orthonormal in the previous chapter. Therefore, the first sum above over n is equal to unity. We may bound the second sum using the above bound on the fourth derivative of $f(x)$:

$$|\sum_{n=0}^{N-1} r_4(\gamma v_k[n]\xi)| \leq \frac{\beta}{4!}\gamma^4\xi^4\lambda_k^2 \sum_{n=0}^{N-1} v_k^4[n] = \frac{\beta}{4!}\gamma^4\xi^4\lambda_k^2 \sum_{n=0}^{N-1} v_k^2[n]v_k^2[n].$$

From Appendix I, $v_k^2[n] \leq (\lambda_k R)^{-1}$. Noting that $|\xi| < a$, and using the orthonormality of the eigenvectors again, we conclude that:

$$|\sum_{n=0}^{N-1} r_4(\gamma v_k[n]\xi)| \leq \frac{\beta\gamma^4 a^4 \lambda_k}{24R},$$

which tends to zero as the oversampling ratio, R, tends to infinity.

Exponentiating Equation 5.9 leads to the result that:

$$\Phi_{\hat{z}[k]}(\xi) = e^{\frac{-\xi^2\gamma^2\lambda_k}{2}} \exp(\sum_{n=0}^{N-1} r_4(\gamma v_k[n]\xi)) = e^{\frac{-\xi^2\gamma^2\lambda_k}{2}} \left(1 + O(R^{-1})\right).$$

Therefore, the characteristic function of the kth eigenmapping converges to $e^{\frac{-\xi^2\gamma^2\lambda_k}{2}}$ *uniformly* for $|\xi| < a$ as the oversampling ratio, $R \to \infty$. Since $e^{\frac{-\xi^2\gamma^2\lambda_k}{2}}$ is the characteristic function for the Gaussian p.d.f. with zero mean and variance $\gamma^2\lambda_k = \frac{\lambda_k c^2}{R}$, the theorem is proved by using Theorem 11 of Cramér [6]. ∎

This marginal p.d.f. result is of interest because it is independent of the relationship between the block size and the oversampling ratio imposed later in this chapter. Therefore, it is valid for any component of the full $N \times 1$ eigenmapping, not just the components in the $N_{VQ} \times 1$ eigenmapping.

5.4 The Joint P.D.F.

Theorem 5.3. *Let* $\frac{N_{VQ}^2}{R} \to 0$ *where* N_{VQ} *is the vector quantizer dimension and* R *is the oversampling ratio. Then the joint p.d.f. of the binary* ($\pm c$) *output eigenmappings tends to a joint Gaussian p.d.f. with zero mean vector and covariance matrix* $\frac{c^2}{R}\Lambda_{N_{VQ}}$.

Comment: Theorem 11a in [6] is the multidimensional extension of Theorem 11 used in Theorem 5.2. In light of this fact, this proof is a multidimensional extension of the proof of Theorem 5.2. In N_{VQ}-dimensional space, the joint characteristic function is the Fourier transform of the joint p.d.f. and is a function of N_{VQ} variables, $\xi_0, \xi_1, \ldots, \xi_{N_{VQ}-1}$. Theorem 11a in [6] states that a necessary and sufficient condition for the convergence of

a sequence of joint p.d.f.'s to a limiting p.d.f. is that the corresponding sequence of joint characteristic functions converge to the limiting joint characteristic function uniformly in every finite multidimensional interval $|\xi_k| < a$, $k = 0, 1, \ldots, N_{VQ} - 1$, for every finite a. Therefore, we consider the joint characteristic function of the binary output eigenmappings as a function of the oversampling ratio, R, in light of the results of Section 5.2.

Proof: The joint characteristic function of the random eigenmapping, $\hat{\mathbf{z}}$, is defined as:

$$\Phi_\xi = \Phi_{\hat{\mathbf{z}}}(\xi_0, \xi_1, \ldots, \xi_{N_{VQ}-1}) = E\{\exp(j \sum_{k=0}^{N_{VQ}-1} \xi_k \hat{z}[k])\} = E\{\exp(j \sum_{n=0}^{N-1} \alpha_\xi[n] z[n])\},$$

where the $\alpha_\xi[n]$ sequence is defined (using Equation 5.7) as:

$$\alpha_\xi[n] = R^{-\frac{1}{2}} \sum_{k=0}^{N_{VQ}-1} \lambda_k^{1/2} \xi_k v_k[n]. \tag{5.10}$$

The ξ subscript denotes dependence on the independent variables $\xi_0, \xi_1, \ldots, \xi_{N_{VQ}-1}$.

Using the definition of the random $z[n]$ sequence as in the previous section, it follows that:

$$\Phi_\xi = \prod_{n=0}^{N-1} \cos(\gamma \alpha_\xi[n]),$$

where we define $\gamma = \frac{c}{R^{1/2}}$. The assumption that γ is constant as $R \to \infty$ is again made and will be verified in a later section. This makes the output scale value, c, a function of the oversampling ratio, R.

The argument of the cosine in the nth product term above can be bounded using Schwarz' inequality on Equation 5.10:

$$|\gamma \alpha_\xi[n]| \leq |\gamma| \sqrt{\sum_{k=0}^{N_{VQ}-1} |\xi_k|^2} \sqrt{\sum_{k=0}^{N_{VQ}-1} \lambda_k v_k^2[n]}.$$

In order to use Theorem 11a in [6] we consider $|\xi_k| < a$ for $k = 0, 1, \ldots, N_{VQ} - 1$ for some a that is arbitrarily large, but fixed. Therefore, the argument of the cosine in each product term can be further bounded using the result of Appendix II of this chapter:

$$|\gamma \alpha_\xi[n]| \leq |\gamma| |a| (\frac{N_{VQ}}{R})^{1/2}, \tag{5.11}$$

which tends to zero when $N_{VQ} \geq 1$ and $\frac{N_{VQ}^2}{R} \rightarrow 0$.

Therefore, the natural logarithm of the joint characteristic function is well-defined as $R \rightarrow \infty$:

$$\ln(\Phi_\xi) = \sum_{n=0}^{N-1} f(\gamma \alpha_\xi[n]),$$

where $f(x) = \ln(\cos(x))$. As in the previous section, we expand $f(x)$ using a Taylor series expansion with a fourth-order remainder, $r_4(x)$:

$$\ln(\Phi_\xi) = -\frac{\gamma^2}{2} \sum_{n=0}^{N-1} \alpha_\xi^2[n] + \sum_{n=0}^{N-1} r_4(\gamma \alpha_\xi[n]). \tag{5.12}$$

Consider the first sum over n above:

$$-\frac{\gamma^2}{2} \sum_{n=0}^{N-1} \alpha_\xi^2[n] = -\frac{\gamma^2}{2} \sum_{n=0}^{N-1} \sum_{k=0}^{N_{VQ}-1} \sum_{l=0}^{N_{VQ}-1} \xi_k \xi_l \lambda_k^{1/2} \lambda_l^{1/2} v_k[n] v_l[n].$$

Interchanging orders of summation, we can evaluate the sum over n from 0 to $N-1$ first. Since the eigenvectors are orthonormal, the quantity $\sum_{n=0}^{N-1} v_k[n] v_l[n]$ is unity when $k = l$ and zero otherwise. Therefore, the first sum over n in Equation 5.12 is:

$$-\frac{\gamma^2}{2} \sum_{n=0}^{N-1} \alpha_\xi^2[n] = -\frac{\gamma^2}{2} \sum_{k=0}^{N_{VQ}-1} \lambda_k \xi_k^2. \tag{5.13}$$

We now bound the second sum over n in Equation 5.12. Recall from the previous section that as the argument of $f(x)$ tends to zero we can bound the fourth derivative by some positive constant β. Therefore, as in the previous section, the second sum over n is:

$$\left| \sum_{n=0}^{N-1} r_4(\gamma \alpha_\xi[n]) \right| \leq \frac{|\beta|}{4!} \sum_{n=0}^{N-1} \gamma^4 \alpha_\xi^4[n].$$

When we express $\gamma^4 \alpha_\xi^4[n] = (\gamma^2 \alpha_\xi^2[n])(\gamma^2 \alpha_\xi^2[n])$, we can use the bound in Equation 5.11 and the equality in Equation 5.13. To use Theorem 11a in [6], we restrict our attention to $|\xi_k| < a$ for $k = 0, 1, \ldots, N_{VQ} - 1$ where a is an arbitrary fixed constant. Recalling from Chapter 4 that the eigenvalues are less than or equal to unity, we may finally write:

$$\left| \sum_{n=0}^{N-1} r_4(\gamma \alpha_\xi[n]) \right| \leq \frac{a^4 \beta \gamma^4 N_{VQ}^2}{24R}.$$

Therefore, as $\frac{N_{VQ}^2}{R} \rightarrow 0$, the second sum over n in Equation 5.12 tends to zero uniformly for all $|\xi_k| < a$ and all finite a. We obtain the joint characteristic function from Equation 5.12 by exponentiating both sides of the equation. This leads to:

$$\Phi_\xi = \exp\left(-\frac{\gamma^2}{2}\sum_{k=0}^{N_{VQ}-1}\lambda_k\xi_k^2\right)\exp(\sum_{n=0}^{N-1} r_4(\gamma\alpha_\xi[n]))$$

$$= \exp\left(-\frac{\gamma^2}{2}\sum_{k=0}^{N_{VQ}-1}\lambda_k\xi_k^2\right)\left(1+O(\frac{N_{VQ}^2}{R})\right).$$

Therefore, we have the desired result that

$$\Phi_\xi \rightarrow \exp\left(-\frac{\gamma^2}{2}\sum_{k=0}^{N_{VQ}-1}\lambda_k\xi_k^2\right)$$

uniformly as $\frac{N_{VQ}^2}{R} \rightarrow 0$ for all $|\xi_k| < a$, for all fixed a. Application of Theorem 11a in [6] completes the proof. ∎

5.5 Experimental Validation

This section presents the outcomes of experiments that were designed to test the validity of joint Gaussian p.d.f. result. While this result requires $\frac{N_{VQ}^2}{R}$ to tend to zero, it is of interest to see how accurate the joint Gaussian prediction is for reasonable values of N_{VQ} and R. The experiments were based on the one-dimensional Kolmogorov-Smirnov (K-S) test [2] which indicates how well a particular cumulative distribution function (c.d.f.) describes a collection of data points. The p.d.f. is the derivative of the c.d.f. For a given collection of scalar data points and a given c.d.f., the one-dimensional K-S test generates a distribution-free measure of the probabilistic validity of the hypothesis that the given collection of data points were obtained from a random process with the given c.d.f. As the number of data points increases, the one-dimensional K-S test becomes more accurate in a manner that is well-understood [2].

The multi-dimensional K-S test used here is a natural generalization of the one-dimensional test. The multi-dimensional K-S test generates a single scalar quantity that is used to reflect the likelihood that a collection of multi-dimensional data points were generated by a random vector process with a particular c.d.f. As opposed to the one-dimensional case, the precise relationship between the scalar quantity generated by the test, the number of data points and the probabilistic validity of the multi-dimensional test is not known [7, 8]. Because of this limitation, it is useful to perform the multi-dimensional K-S test on a "control" set of data points for purposes of comparison. The "control" set used here will be a collection of data points obtained from a pseudo-random process that has the desired c.d.f. In the case at hand, the desired c.d.f. is Gaussian with zero mean vector and covariance matrix $\frac{c^2}{R}\Lambda_{N_{VQ}}$.

We now define the multi-dimensional K-S test. Let $\mathcal{E}_{N_P}$ be a set containing N_P N_{VQ}-dimensional points. These can be viewed as the data points obtained from some experiment that we wish to compare with a given distribution. Based on this collection of points, we can define another set, $\mathcal{S}_{\mathcal{E}}(\hat{\mathbf{x}})$, that contains all the points in $\mathcal{E}_{N_P}$ that are "below" the point $\hat{\mathbf{x}}$:

$$\mathcal{S}_{\mathcal{E}}(\hat{\mathbf{x}}) = \{\hat{\mathbf{z}} \in \mathcal{E}_{N_P} : \hat{z}[k] \leq \hat{x}[k], k = 0, 1, \ldots, N_{VQ} - 1\}.$$

Based on this collection of points, we can define a function $P_{\mathcal{E}}(\hat{\mathbf{x}})$ to be the fraction of points in $\mathcal{E}_{N_P}$ "below" the point with coordinates $\hat{\mathbf{x}}$:

$$P_{\mathcal{E}}(\hat{\mathbf{x}}) = \frac{|\mathcal{S}_{\mathcal{E}}(\hat{\mathbf{x}})|}{N_P}.$$

This function can be viewed as an estimate of the c.d.f. of the process that generated the given data points. Finally, we define the metric that is the output of this multi-dimensional K-S test:

$$D_{MAX} = \max_{\hat{\mathbf{x}} \in \mathcal{R}^{N_{VQ}}} |P_{\mathcal{E}}(\hat{\mathbf{x}}) - P(\hat{\mathbf{x}})|$$

where $P(\hat{\mathbf{x}})$ is the c.d.f. that we wish to test the given points against.

Three types of experiments were performed. The first type of experiment operated on the "control" data set discussed above. The next two types of experiments operated on a subset of the total number of output eigenmappings of single-bit oversampled data converters with

oversampling ratios of $R = 10$ and $R = 100$. A subset of size N_P was randomly chosen because an exhaustive survey of all 2^N binary output eigenmappings is more than a little prohibitive when $N \geq 100$, as will be the case here. In these last two types of experiments, the output scale value, c, was selected so that $\frac{c^2}{R} = 1$ and each point in the data set was obtained by taking the eigenmapping of a pseudo-randomly chosen binary ($\pm c$) vector of length $N = N_{VQ}R$. This selection of the output scale value will be motivated in Section 5.6. Each type of experiment consisted of seven K-S tests performed on $N_P = 100,000$ data points. The dimension of the vector quantizer in each test was $N_{VQ} = 10$. The K-S metrics presented in each entry of the following table were averaged over all seven tests. Standard deviations are also given.

Experiment	K-S Metrics: Mean $\pm$ Std. Dev.
Control	$(2.30 \pm 0.44) \times 10^{-3}$
R=10	$(4.07 \pm 0.88) \times 10^{-3}$
R=100	$(2.43 \pm 0.51) \times 10^{-3}$

These test results are consistent with the joint p.d.f. result in Theorem 5.3. Using the control experiment as a reference, these experiments suggest that the joint p.d.f. when $R = 100$ is more Gaussian than the joint p.d.f. when $R = 10$, as predicted by the theory. Also, the K-S metrics for the control experiments and for the $R = 100$ experiments are within standard deviations of one another. For both values of R, the K-S metrics are the same order of magnitude as the K-S metric in the control case. These results suggest that the joint Gaussian p.d.f. result above will be valid for reasonable values of oversampling ratio and vector quantization dimension.

5.6 On the Asymptotically Optimal Output Scaling for a Gaussian Input Process

This section presents the three known necessary conditions for an optimum vector quantizer given the joint p.d.f. of the input vector process and a distortion metric [1]. Every vector

quantizer is defined by its mapping function (i.e., the rule that selects an output vector based on an input vector) and its output vectors. The three known necessary conditions for optimal vector quantizer design are:

Mapping Condition The mapping function should be as follows: if the input vector lies in the kth nearest-neighbor region (defined in Section 5.2), then the kth output vector should be chosen to minimize the output distortion.

Boundary Condition The probability that the input vector lies on the boundary of any nearest-neighbor region should be zero.

Centroid Condition The kth output vector should be the centroid of the kth nearest-neighbor region with respect to the joint p.d.f. of the input vector process.

As noted in Section 5.2, the mapping condition should always be satisfied in eigenmodulators that minimize the output distortion. The boundary condition is satisfied in the case of an input with a continuous p.d.f. The approximation of a continuous p.d.f. is valid for discretely distributed input when the number of bits used to represent the input becomes large.

The centroid condition yields the most interesting information about the single-bit eigenmodulator when $\frac{N_{VQ}^2}{R} \to 0$ and the input vector process is joint Gaussian as in Appendix III. According to Appendix III and Theorem 5.3, both the input and output eigenmappings have joint Gaussian distributions with zero mean vectors in this asymptotic scenario. The input eigenmapping has covariance matrix $\sigma^2 \Lambda_{N_{VQ}}^2$ and the output eigenmapping has covariance matrix $\frac{c^2}{R}\Lambda_{N_{VQ}}$, where $\Lambda_{N_{VQ}}$ is the diagonal matrix containing the first N_{VQ} eigenvalues of the filter matrix, $\mathbf{A}$, defined in Chapter 4.

Let N_{VQ} be a function of the oversampling ratio, R, such that $N_{VQ} \to \infty$ and $\frac{N_{VQ}^2}{R} \to 0$. For arbitrarily small $\epsilon > 0$, there exists a block size, N, such that $N_{VQ} = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$. This is because we can make $N = \frac{N_{VQ}R}{1+\epsilon}$, which will be a function of the oversampling ratio also. The block size still tends to infinity, but it does so at a rate that is coupled to the oversampling ratio. In contrast, the block size of the asymptotically equivalent vector quantization problem

described in Theorem 4.4 tended to infinity at a rate independent of the oversampling ratio. The block size is presently chosen so that the number of vector quantization dimensions, N_{VQ}, has the same form as it did for the asymptotically equivalent vector quantization problem in the previous chapter. Recall that Theorem 4.4 expresses the distortion metric of the oversampled data converter problem as the sum of a vector quantization distortion and a residual distortion. Combining Theorems 4.3 and 4.4, the residual distortion term is $O(\exp(-\frac{\alpha N}{R}))$, for some positive constant α. In Theorem 4.4, we could always ensure that the residual distortion was less than any positive quantity by making N large enough for a given R. We cannot presently claim that this vector quantization problem is asymptotically equivalent to the oversampled data converter problem since the residual distortion cannot be made arbitrarily small for a given R, since N is a function of R. For $N_{VQ} = R^{\frac{1}{2}-\eta}$ where $0 < \eta < \frac{1}{2}$, it follows that the residual distortion term is $O(\exp(-\alpha R^{\frac{1}{2}-\eta}))$, where α is some positive constant.

At present we do not know what the vector quantization distortion in the eigenmodulator will be as a function of the oversampling ratio; this is an open research problem. However, we may estimate it by considering modern oversampled data converters. The Synthesis architecture in Chapter 3 of this thesis and the popular $\Delta\Sigma$ modulators [9] have distortion that is inversely polynomial in the oversampling ratio. The residual distortion in the above asymptotic scenario is $O(\exp(-\alpha R^{\frac{1}{2}-\eta}))$ and is therefore much smaller than the total oversampled data converter distortion when this is inversely polynomial in R. We therefore conjecture that the vector quantization problem in the asymptotic scenario of this chapter, $N_{VQ} \to \infty$ and $\frac{N_{VQ}^2}{R} \to 0$, is asymptotically equivalent to the oversampled data converter problem described in Chapter 3. If this conjecture is incorrect then the following discussion regarding the optimal output scaling is invalid, but then the vector quantization distortion would be competitive with $O(\exp(-\alpha R^{\frac{1}{2}-\eta}))$ which would be a breakthrough result by modern oversampled data converter standards.

We wish to compare the covariance matrices $\sigma^2\Lambda_{N_{VQ}}^2$ and $\frac{c^2}{R}\Lambda_{N_{VQ}}$. Consider the N_{VQ} eigenvalues involved: $\lambda_0, \lambda_1, \ldots, \lambda_{N_{VQ}-1}$. Define $N_{VQ2} = \lfloor \frac{N}{R}(1-\epsilon) \rfloor$, where $N_{VQ} = \lfloor \frac{N}{R}(1+\epsilon) \rfloor$.

As seen in Theorem 4.3, the first N_{VQ2} eigenvalues tend to unity at a rate such that $1-\lambda_k = O(\exp(-\frac{\alpha N}{R}))$. Therefore, $\lambda_k^2 \to \lambda_k \to 1$ for $0 \leq k \leq N_{VQ2}$.

We cannot make the same statement for all of the remaining $N_{VQ} - N_{VQ2}$ eigenvalues. But these represent an arbitrarily small fraction of all of the eigenvalues since

$$\frac{N_{VQ} - N_{VQ2}}{N_{VQ}} \sim 2\epsilon,$$

where we are free to make ϵ as small as we like. Therefore, almost all of the diagonal elements of the covariance matrices $\sigma^2\Lambda_{N_{VQ}}^2$ and $\frac{c^2}{R}\Lambda_{N_{VQ}}$ are asymptotically equal to each other when we choose the output scale value to be:

$$c = \sigma R^{1/2},$$

as the oversampling ratio, R, tends to infinity. This suggests that the above scale factor is a good choice for the present asymptotic scenario. In the oversampled data converter literature [9], the output scale factor is typically chosen large to avoid no-overload conditions and so that the output quantizer can be more accurately modeled as an additive white noise source uniformly distributed over a quantization interval.

Appendix I: A Bound on the Eigenvector Components

Lemma A.1. *For all $0 \leq k < N$, all $0 \leq n < N$ and all $R \geq 1$:*

$$v_k[n] \leq (\lambda_k R)^{-1/2}. \qquad (A.1.1)$$

Proof:

The characteristic equation for the eigenvectors, $\mathbf{v}_k$, and eigenvalues, λ_k, of the $N \times N$ filter matrix $\mathbf{A}$ is:

$$\lambda_k \mathbf{v}_k = \mathbf{A}\mathbf{v}_k, \qquad (A.1.2)$$

where $k = 0, 1, \ldots, N-1$. The eigenvalues and eigenvectors are ordered so that $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{N-1}$. Using the definition of the filter matrix in Equation 4.3 of the previous

chapter, the characteristic equation may be written using a frequency domain integral as:

$$\lambda_k v_k[n] = \int_{-\pi/R}^{\pi/R} V_k(e^{jw}) e^{jwn} \frac{d\omega}{2\pi}, \qquad (A.1.3)$$

where $v_k[n]$ is the nth component ($n = 0, 1, ..., N-1$) of the kth eigenvector, $\mathbf{v}_k$, and $V_k(e^{jw})$ is defined as its Fourier transform:

$$V_k(e^{jw}) = \sum_{n=0}^{N-1} v_k[n] e^{-jwn}. \qquad (A.1.4)$$

It was shown in the previous chapter that the eigenvalues are non-zero. Using Schwarz's inequality, we may bound the magnitude of $v_k[n]$ by:

$$|v_k[n]| \leq \frac{1}{\lambda_k} \left(\int_{-\pi/R}^{\pi/R} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi} \right)^{1/2} \left(\int_{-\pi/R}^{\pi/R} \frac{d\omega}{2\pi} \right)^{1/2}. \qquad (A.1.5)$$

The second integral on the right-hand side above is

$$\int_{-\pi/R}^{\pi/R} \frac{d\omega}{2\pi} = \frac{1}{R}.$$

We now wish to evaluate the first integral on right-hand side of Equation A.1.5. Multiply both sides of Equation A.1.3 by $v_k[n]$ and sum from $n = 0$ to $n = N - 1$. Using the fact from the previous chapter that the eigenvectors are orthonormal, we obtain:

$$\lambda_k = \int_{-\pi/R}^{\pi/R} V_k(e^{jw}) \left(\sum_{n=0}^{N-1} v_k[n] e^{jwn} \right) \frac{d\omega}{2\pi} = \int_{-\pi/R}^{\pi/R} |V_k(e^{jw})|^2 \frac{d\omega}{2\pi}.$$

Therefore, the first integral on the right-hand side of Equation A.1.5 equals λ_k and completes the proof. ∎

Appendix II: A Useful Bound on a Sum of Scaled, Squared Eigenvectors

Lemma A.2. *For all $0 \leq N_{VQ} < N$, all $0 \leq n < N$ and all $R \geq 1$:*

$$\sum_{k=0}^{N_{VQ}-1} \lambda_k v_k^2[n] \leq \frac{1}{R}. \qquad (A.2.1)$$

Proof: Since for all $k = 0, 1, \ldots, N-1$, $\lambda_k v_k^2[n]$ is a non-negative quantity, we have the preliminary inequality:

$$\sum_{k=0}^{N_{VQ}-1} \lambda_k v_k^2[n] \leq \sum_{k=0}^{N-1} \lambda_k v_k^2[n]. \qquad (A.2.2)$$

We now evaluate the expression on the right-hand side. From the characteristic equation for the eigenvectors defined in the previous chapter,

$$\lambda_k v_k[n] = \sum_{m=0}^{N-1} \mathbf{A}_{n,m} v_k[m], \qquad (A.2.3)$$

where $\mathbf{A}_{n,m} = \frac{1}{R}\text{sinc}(\frac{n-m}{R})$. When we multiply both sides of the characteristic equation by $v_k[n]$ and sum from $k = 0$ to $k = N-1$ we obtain:

$$\sum_{k=0}^{N-1} \lambda_k v_k^2[n] = \sum_{m=0}^{N-1} \mathbf{A}_{n,m} \sum_{k=0}^{N-1} v_k[n] v_k[m]. \qquad (A.2.4)$$

Recall that $v_k[n]$ is the (n,k) component of the unitary matrix $\mathbf{V}$ that diagonalizes $\mathbf{A}$. Because the matrix $\mathbf{V}$ is unitary, its rows are orthonormal as well as its columns [10]. Therefore,

$$\sum_{k=0}^{N-1} v_k[n] v_k[m] = \delta[n-m] = \begin{cases} 1 & n = m \\ 0 & \text{otherwise} \end{cases} \qquad (A.2.5)$$

Therefore, the sum over m in Equation A.2.4 collapses to the single $n = m$ term, $\mathbf{A}_{n,n} = \frac{1}{R}$. The proof is completed by combining the preliminary inequality in Equation A.2.2 with this result. ∎

Appendix III: The Eigenmapping of a Gaussian Input Vector

It is of interest to study the eigenmapping of a vector formed from samples of a Gaussian process with variance σ^2 that is band-limited to the frequency interval $(-\frac{\pi}{R}, \frac{\pi}{R})$. This random input vector, $\mathbf{x}$, is described by a joint Gaussian p.d.f. with zero mean vector and covariance matrix $\Phi_{\mathbf{x}} = E\{\mathbf{x}\mathbf{x}^\dagger\} = \sigma^2 R\mathbf{A}$. The importance of the filter matrix, $\mathbf{A}$, was detailed in the previous chapter. The filter matrix may be written as $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\dagger$, where $\Lambda =$

$\text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$ is the diagonal matrix containing the N eigenvalues of $\mathbf{A}$ and $\mathbf{V}$ is the $N \times N$ unitary matrix that diagonalizes $\mathbf{A}$. Consider the following transformation:

$$\mathbf{x}' = R^{-\frac{1}{2}} \Lambda^{1/2} \mathbf{V}^\dagger \mathbf{x}.$$

Since this is a non-singular linear transformation, the new vector, $\mathbf{x}'$, is also joint Gaussian [3] with zero mean vector and covariance matrix:

$$\Phi_{\mathbf{x}'} = E\{\mathbf{x}'(\mathbf{x}')^\dagger\} = R^{-1} \Lambda^{1/2} \mathbf{V}^\dagger E\{\mathbf{x}\mathbf{x}^\dagger\} \mathbf{V} \Lambda^{1/2} = \sigma^2 \Lambda^2.$$

Since the covariance matrix is diagonal, the components of this transformed vector are uncorrelated with respect to each other. Further, the components are independent of each other because they are uncorrelated and joint Gaussian [3]. We now use this information to analyze the $N_{VQ} \times 1$ eigenmapping of the input vector.

For $1 \leq N_{VQ} < N$, the $N_{VQ} \times 1$ eigenmapping of the input vector was defined in the previous chapter to be:

$$\hat{\mathbf{x}} = R^{-\frac{1}{2}} \Lambda_{N_{VQ}}^{1/2} \mathbf{V}_{N_{VQ}}^\dagger \mathbf{x},$$

where $\Lambda_{N_{VQ}} = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N_{VQ}-1})$ and $\mathbf{V}_{N_{VQ}}$ is the $N \times N_{VQ}$ matrix formed from the first N_{VQ} columns of $\mathbf{V}$. Since the kth component of $\hat{\mathbf{x}}$ equals the kth component of $\mathbf{x}'$ for all $0 \leq k < N_{VQ}$, we may integrate the joint p.d.f. of $\mathbf{x}'$ over the appropriate region of space to show that the $N_{VQ} \times 1$ eigenmapping has a p.d.f. that is joint Gaussian with zero mean vector and covariance matrix $\sigma^2 \Lambda_{N_{VQ}}^2$. We have thus proved the following lemma:

Lemma A.3. *Let an $N \times 1$ random input vector be described by a joint Gaussian p.d.f. with zero mean vector and covariance matrix $\sigma^2 R\mathbf{A}$, where $\mathbf{A}$ is the filter matrix described in the previous chapter. For $1 \leq N_{VQ} \leq N$, its $N_{VQ} \times 1$ eigenmapping is a random vector described by a joint Gaussian p.d.f with zero mean vector and covariance matrix $\sigma^2 \Lambda_{N_{VQ}}^2$, where $\Lambda_{N_{VQ}}$ is the diagonal matrix containing the first N_{VQ} eigenvalues of the filter matrix.*

References

1. A. Gersho and R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, 1991.

2. R. von Mises, Mathematical Theory of Probability and Statistics, Academic Press, New York, 1964.

3. W. Feller, An Introduction to Probability Theory and Its Applications, Volume II, Wiley, New York, 1966.

4. A. Gersho, "Asymptotically Optimal Block Quantization," *IEEE Transactions on Information Theory,* vol. IT-25, pp. 373-380, July 1979.

5. Y. Yamada, S. Tazaki, R. Gray, "Asymptotic Performance of Block Quantizers with Difference Distortion Measures," *IEEE Transactions on Information Theory,* vol. IT-26, pp. 6-14, January 1980.

6. H. Cramér,Random Variables and Probability Distributions, Cambridge University Press, London, 1937.

7. J. Kiefer, "On Large Deviations of the Empiric Distribution Function of Vector Chance Variables and a Law of the Iterated Logarithm," Pac. Math. 11, pp. 649-660, 1961.

8. J. Kiefer, J. Wolfowitz, "On the Deviations of the Empiric Distribution Function of Vector Chance Variables," Trans. Amer. Math. Soc. 87, 173-186, 1958.

9. *Oversampled Data Converters: Theory, Design and Simulation,* edited by J. Candy, G. Temes, IEEE Press, New York, 1992.

10. J. Franklin, Matrix Theory, Prentice-Hall, Englewood Cliffs, NJ, 1968.

Chapter 6

Future Work

6.1 Digital Sinusoid Generators

Future work for the digital sinusoid generators has two aspects: improved dither signals and highly-linear digital-to-analog converters (DACs). It is of interest to find dither signals that give good spurious performance and yet are not spectrally white but instead have band-pass characteristics. In phase dithering, the total quantization noise is amplitude modulated onto the fundamental. If this noise is high-pass, then much of it could be rejected by either an analog or digital phase-locked loop filter, depending on the application. In amplitude dithering, the total quantization noise is additive. If the total quantization noise could be minimized over some frequency band, then output frequencies generated in this band would have superior local noise performance compared to the dithering systems in Chapter 2.

Because of amplitude and phase dithering, the spurious performance of analog systems employing digital sinusoid generators is no longer tied to the number of bits sent to the DAC. However, the DAC non-linearities can give rise to large spurs even when the digital input has virtually no spurs. The manufacturers of DACs for these digital frequency synthesis applications need to divorce linearity performance from the DAC input resolution. As implied by the dual nature of this thesis, highly-linear oversampled data converters may someday be good candidates for DACs in DDS applications.

6.2 Oversampled Data Converters

For the synthesis approach, it is of interest to consider architectures that result from selecting two or more output samples at a time in order to minimize an appropriate distortion metric. Other distortion metrics, including those that measure spurious harmonic performance, should be investigated. Additional tests with other types of input signals (especially sinusoids) should also be performed.

The construction of eigenmodulators is an important item for future work. Their absence at present is due to the large complexity of the vector quantizers (VQs) at their heart. With the appropriate VQ technological advances, eigenmodulators will become increasingly feasible. At present, they have theoretical potential. My present eigenmodulator research focuses on the evaluation of the vector quantization distortion under the asymptotic scenarios presented in Chapter 5. The vector quantization distortion computations are facilitated by information about the asymptotic distribution of the output eigenmappings. Therefore, it is also of interest to determine the distribution of the output eigenmappings when the block size grows in a manner uncoupled to the oversampling ratio. Distribution information will also undoubtedly prove useful in the design of reduced-complexity vector quantizers in the eigenmodulators.