Optimal Dither and Noise Shaping in Image Processing

by

Cameron Nicklaus Christou

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Dithered quantization and noise shaping is well known in the audio community. The image processing community seems to be aware of this same theory only in bits and pieces, and frequently under conflicting terminology. This thesis attempts to show that dithered quantization of images is an extension of dithered quantization of audio signals to higher dimensions.

Dithered quantization, or “threshold modulation”, is investigated as a means of suppressing undesirable visual artifacts during the digital quantization, or requantization, of an image. Special attention is given to the statistical moments of the resulting error signal. Afterwards, noise shaping, or “error diffusion” methods are considered to try to improve on the dithered quantization technique.

We also take time to develop the minimum-phase property for two-dimensional systems. This leads to a natural extension of Jensen’s Inequality and the Hilbert transform relationship between the log-magnitude and phase of a two-dimensional system. We then describe how these developments are relevant to image processing.

Acknowledgements

I would like to thank my supervisors, Dr. Stanley Lipshitz and Dr. Bernhard Bodmann, for their support and the opportunity to pursue this work. Their encouragement and enthusiasm has been greatly appreciated over the course of my research.

I would also like to thank Dr. John Vanderkooy, Kevin Krauel, and Jim Hayward for many entertaining and enlightening conversations along the way. A special thanks to all the graduate students I have worked with over the years as well, most notably: Eddie DuPont, Colin Turner, Rudy Gunawan, Kathleen Wilkie, Sean Speziale, Michael LaCroix, Lalit Jain, Mohamed Omar, and Ben Smith. You have all made this academic experience a memorable one.

I would like to thank the entire Applied Math department at the University of Waterloo for the time and effort spent teaching me over the years.

Finally I would like to thank Dr. John Wainwright and Kristine Melloh for helping me to make it this far.

This is dedicated
to my parents, Nick and Pam.
With love.

Contents

List of Figures

Introduction

The problem we are concerned with in this thesis is that of quantization of signals. Quantization is a procedure which necessarily causes a degradation of the original signal, but is also necessary in order to record the signal onto digital media. The two applications considered in this thesis are digital audio and digital image processing. The purpose of this thesis is to demonstrate that the distortions caused by quantization can not be eliminated, but they can be manipulated so that the original signal is qualitatively preserved after the quantization process.

Dithered quantization and noise shaping are techniques that are widely used in the field of digital audio processing today. Research into the subject has shown that there has been an interest of applying similar techniques to digital image processing, but that the theory developed on this topic is relatively incomplete by comparison. The image processing community appears to be aware of dithered quantization, or "threshold modulation", but does not seem to be aware of how different dither signals can affect the quantized signal. In this thesis, we consider a statistical approach to the quantized output and consider the statistical moments of the distortion caused by the quantization process. The image processing community is aware of the concept of noise shaping, or "error diffusion", but does not consider using dither and noise shaping together.

Along with the interest in noise shaping of images comes a desire to create a noise shaper with a given log-magnitude response. As a result, there is some literature on the topic of finding a minimum-phase system with a given log-magnitude with respect to finding a system with a stable inverse, however there has been no mention of why one would desire the minimum-phase system and why no other system will perform just as well for applications in image processing. The optimality of the minimum-phase system for applications in noise shaping is made clear in this thesis. Furthermore, the design methods presented throughout the literature for creating a minimum-phase system have all been met with counterexamples. This thesis offers a design method for finding the minimum-phase system associated with a given log-magntiude. We prove that this method actually produces the minimum-phase system associated with a given log-magnitude, and demonstrate its validity by testing it on the counterexamples found for other methods in the literature.

This thesis will bridge the gap between the two disciplines, and show that dithered quantization and noise shaping is applicable to signals of arbitrary finite dimensions.

Chapter 1

Dither and Noise Shaping in Audio

Before we can study dither and quantization of a real signal of two independent variables, $f(x_1, x_2)$, we review the theory as it pertains to signals of one independent variable, $f(t)$. We will build up this theory with a view toward digital audio processing. The full theory is not be developed here, since this theory is well understood and discussed at length in other works [60, 27, 59, 6, 63, 81, 83, 73, 62, 82]. Instead, we present the salient features of the theory so that the reader has an idea of the concepts we are extending in later chapters. References are provided for the reader who wishes a deeper understanding of the concepts discussed in this chapter.

We begin by explaining the process of sampling and quantization of a continuous signal, $f(t)$. We then discuss how dithered quantization can improve the perceptual quality of the recreated signal after sampling and quantization. This chapter concludes with a discussion of noise-shaping techniques and the uses and design of minimum-phase systems.

Throughout this chapter, sums are assumed to be over all integer values of the index unless otherwise specified.

1.1 Sampling and Quantization

The theory of digital signal processing starts with the Sampling Theorem. The theorem itself tells us how we can transform a continuous signal into discrete packets of data without losing any information.

Theorem 1.1.1 (Sampling Theorem for Trigonometric Polynomials [31]). *Given a trigonometric polynomial with period T,*

$$f(t) = \sum_{m=-N}^{N} c_m e^{i2\pi m \frac{t}{T}},$$

we can reconstruct the signal $f(t)$ exactly from equally spaced sample points during one period of *the signal, as long as we sample the function at least $2N+1$ times. If P is odd and $P \geq 2N+1$, then we have the explicit reconstruction formula*

$$f(t) = \sum_{q=0}^{P-1} f\left(q\frac{T}{P}\right) \frac{\sin(\pi(t\frac{P}{T} - q))}{P\sin(\frac{\pi}{P}(t\frac{P}{T} - q))}.$$

If P is even and $P \geq 2N+1$, then we have the explicit reconstruction formula

$$f(t) = \sum_{q=0}^{P-1} f\left(q\frac{T}{P}\right) \left(\frac{\sin(\pi(t\frac{P-1}{T} - q\frac{P-1}{P}))}{P\sin(\frac{\pi}{P}(t\frac{P}{T} - q))} + \frac{1}{P}\cos\left(\pi\left(t\frac{P}{T} - q\right)\right)\right).$$

The original Sampling Theorem can be traced back to Whittaker [87], and has many generalizations. The form used in this thesis is not the most general, but is the most suitable for the specific purposes of this thesis. It can be derived directly from the original sampling theorem by assuming that the function we are sampling is a trigonometric polynomial. We use this form since we are often talking about sampling a finite sequence.

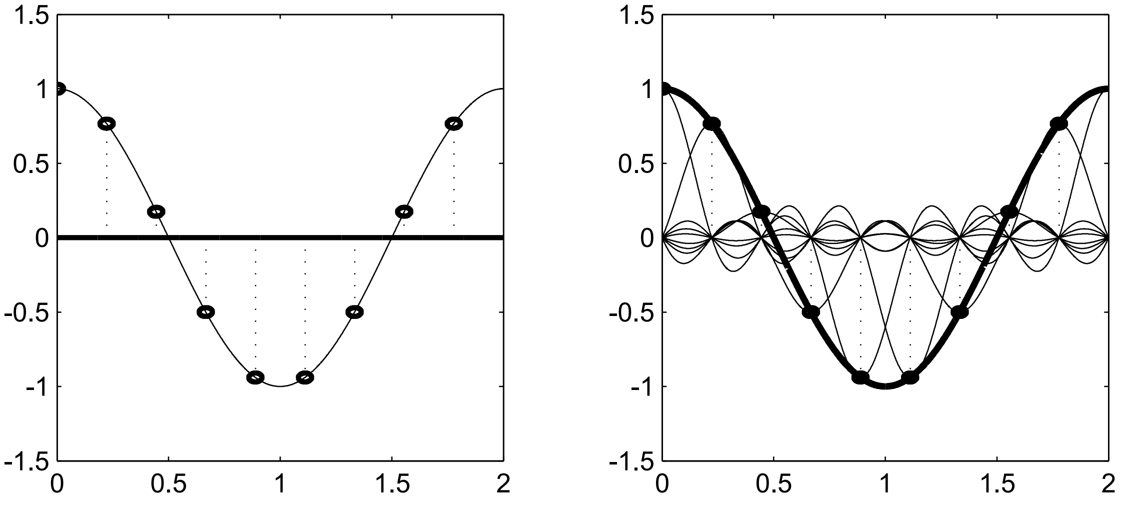
When working with a finite signal, such as a sound clip or an image, we will assume that the signal is extended periodically and that this portion of the signal represents one period. We do this to match the form of the Theorem 1.1.1.

Example 1.1.2. Consider the function $\cos(\pi t)$. For this example, we will sample the function every $\frac{2}{9}$ seconds, giving us $P = 9$ and $T = 2$. As we see in Figure 1.1, the function is reproduced exactly from the sample points as a sum of interpolating functions. The figure shows one period of the original signal.

♣

The sampling frequency is given by $f_s = \frac{P}{T}$. The Nyquist frequency is defined to be $\frac{P}{2T}$. Both of these quantities are usually converted to familiar units, such as Hertz, in applications. If the signal $f(t)$ is not sampled at a rate greater than the Nyquist rate $\frac{2N}{T}$ (i.e. $P < 2N+1$), then the reconstructed signal will be falsified by artifacts known as "aliases".

The sampling theorem is essential to the concept of digital data. It tells us that we can reconstruct a signal perfectly as long as we sample it fast enough. However, the sampling theorem requires infinite precision of the sampled values, $f\left(\frac{k}{f_s}\right)$, and the time of sampling. In order to digitally transmit or store these sample values, we are forced to truncate or round the sample values to finite precision in a process called quantization. Thus we are left with slightly distorted sample values with which to reconstruct our signal. This inevitably introduces a signal dependent error. In the case of digital audio, this error is apparent as an undesirable distortion of the intended sound.



(a)

(b)

Figure 1.1: a) Original cosine signal. b) Cosine signal reconstructed from sample points.

The sample points are ideally rounded, or truncated, to a set of values that are uniformly spaced. This procedure is done by a quantizer. The distance between these quantization levels will be referred to as the quantizer step size Δ, or Least Significant Bit (LSB). There are two different quantizers that are generally used. The first is the midtread quantizer

$$Q(x) = \Delta \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor .$$

The second is the midriser quantizer

$$Q(x) = \Delta \left\lfloor \frac{x}{\Delta} \right\rfloor + \frac{\Delta}{2}.$$

The names of these quantizers are based on the locations of the origin in their graphs. See Figure 1.2. Note that $\lfloor x \rfloor$ denotes the greatest integer value that is less than or equal to x.

The quantization error will be defined as

$$q(x) = Q(x) - x.$$

Example 1.1.3. Consider $\cos(\pi t)$ once again. Suppose that the sample points obtained in Example 1.1.2 had to be quantized to the nearest integer multiple of 0.5. This changes the shape of the reconstructed signal. See Figure 1.3. The reconstructed signal displays a signal dependent error, as shown in Figure 1.4.

♣

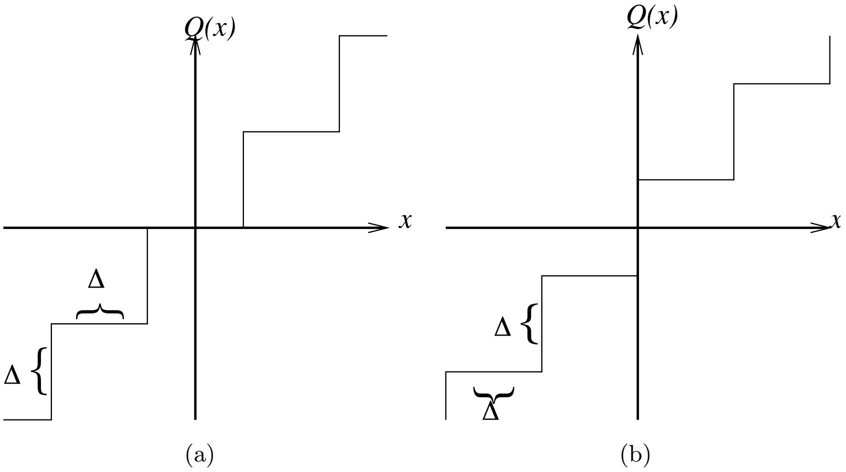


Figure 1.2: a) Midtread Quantizer. b) Midriser Quantizer.

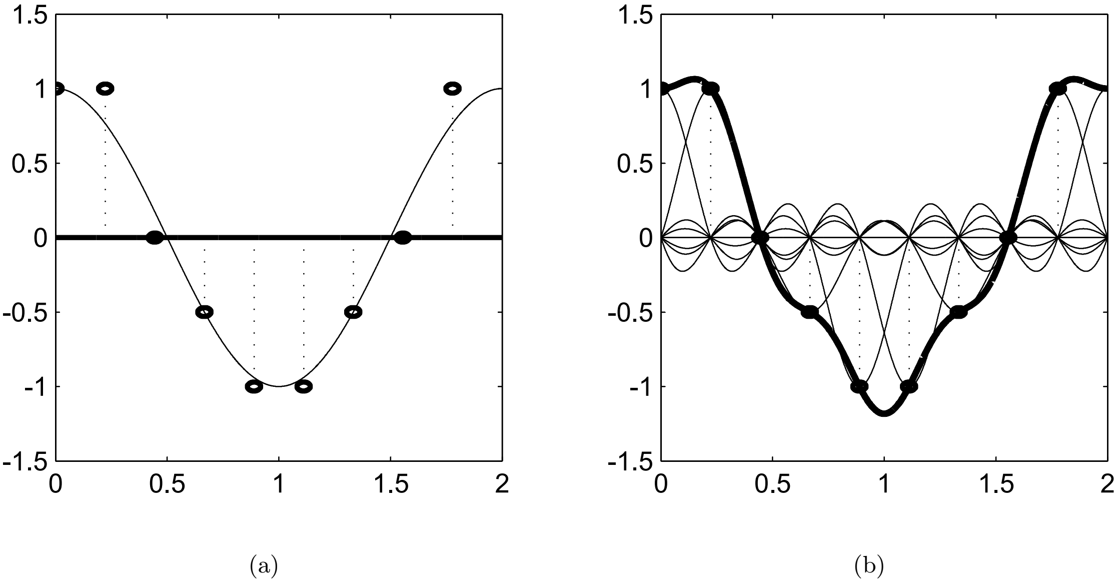


Figure 1.3: a) Quantized cosine signal. b) Cosine signal reconstructed from quantized points.

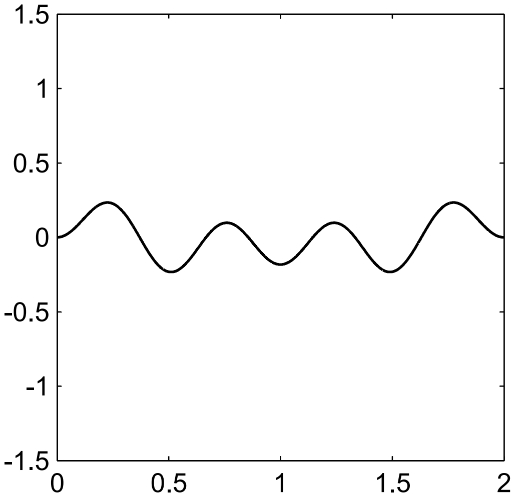


Figure 1.4: Error of the reconstructed signal after quantization.

In theory, these quantizers are assumed to be of infinite capacity. In practice, however, a quantizer can be saturated. That is, the quantizer will only perform properly as long as the input signal is not too large in amplitude. If the input signal becomes too large for the quantizer to handle, then all of the following statistical analysis of dithered quantization will be invalid.

In order to understand the theory of dithered quantizers, it is important to understand what is happening in the undithered case. Figure 1.5 shows the basic undithered quantization schematic. For the undithered system, the error, $\epsilon = Q(x) - x$, is equal to the quantization error, q. The quantity ϵ will be referred to as the total error of the dithered quantizer. For this case, the dither is a zero signal. The reason for this terminology will become apparent later.

Throughout this development, we will consider the input signal to be a random trigonometric polynomial. When we talk about inputting a sequence into a system, we are referring to a sampled random trigonometric polynomial. Since we can not assume that the input has a particular form, this allows us to develop a statistical analysis of the system. The function spaces are finite dimensional, so there are no technical problems caused by this step.

Definition 1.1.4. Let $p(x)$ be the probability density function (pdf) of a random variable. The characteristic function of this random variable is

$$P_x(u) = \int_{-\infty}^{\infty} p(x) e^{i2\pi ux} dx.$$

In other words, the characteristic function of a random variable is the inverse Fourier transform of its pdf. ⧫

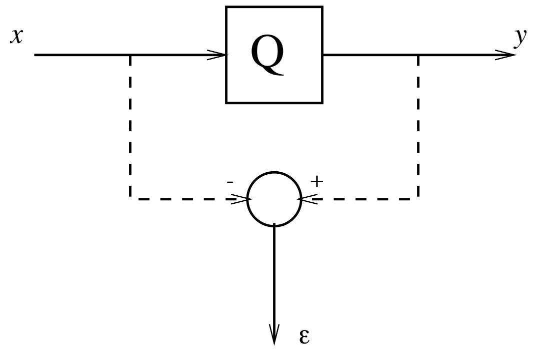


Figure 1.5: Schematic of the undithered quantizer.

Theorem 1.1.5 ([75]). *The pdf of the total error of an undithered quantizing system is uniformly distributed and of width Δ if and only if the characteristic function of the system input, P_x, satisfies $P_x(u) = 0$ for $u = \pm\frac{1}{\Delta}, \pm\frac{2}{\Delta}, \ldots$.*

What this theorem tells us is that if an input signal to an undithered quantizer has certain statistical properties, then the system output will be the original input signal with an added error signal of zero mean and constant variance [75]. In fact, it can be shown that all statistical moments of the total error obey the following formula [73]:

$$E[\epsilon^m] = \begin{cases} \dfrac{1}{m+1}\left(\dfrac{\Delta}{2}\right)^m, & m \text{ is even} \\ 0, & \text{otherwise} \end{cases} . \tag{1.1}$$

This does not imply that the error is independent of the input signal. Moreover, this theorem does not guarantee that the error is spectrally white. The term “spectrally white” means roughly that the error signal is composed of equal, non-zero contributions from all frequencies. That is, the Fourier transform of such an error signal is constant in magnitude.

In order to ensure that the error signal is spectrally white, we require the following theorem.

Theorem 1.1.6 ([75]). *In an undithered quantizing system, the joint probability density function $p_{\epsilon_1,\epsilon_2}$ of the total error values ϵ_1 and ϵ_2, separated in time by $\tau \neq 0$, is given by*

$$p_{\epsilon_1,\epsilon_2}(\epsilon_1, \epsilon_2) = \Pi_\Delta(\epsilon_1)\Pi_\Delta(\epsilon_2)$$

if and only if the joint characteristic function P_{x_1,x_2} of the corresponding system inputs x_1 and x_2 satisfies the condition that

$$P_{x_1,x_2}(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}) = 0$$

for $k_1 = 0, \pm1, \pm2, \ldots$, and $k_2 = 0, \pm1, \pm2, \ldots$, but not $(k_1, k_2) = (0, 0)$.

The function Π_Δ is defined as

$$\Pi_\Delta(x) = \begin{cases} \frac{1}{\Delta}, & \frac{-\Delta}{2} < x < \frac{\Delta}{2} \\ \frac{1}{2\Delta}, & |x| = \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases} .$$

This theorem ensures that the total error of the undithered quantizer is spectrally white. If the conditions of Theorem 1.1.6 are satisfied, then Theorem 1.1.5 is also satisfied.

Unfortunately, many of the signals that need to be quantized do not satisfy the conditions of Theorem 1.1.6. Quantizing a signal such as the one chosen in Example 1.1.3 will result in the original signal with a signal dependent distortion. The fact that the error is recognizably dependent on the original signal is the reason why it is audibly undesirable. Other input signals also display other undesirable artifacts, which we will discuss later.

Example 1.1.7. Consider a 1 second clip of a 1kHz sine wave of amplitude 2Δ, as seen in Figure 1.6(a), sampled at 44.1kHz (as on an audio compact disc) and quantized to the nearest Δ, shown in Figure 1.6(b). In particular, Figure 1.6(c) shows the power spectral density (or spectrum, magnitude of the Fourier transform) of the quantized signal. The power spectral density gives us an indication of the frequency content of a signal. The spectrum of this signal has a dominant peak at 1kHz, telling us that the signal behaves largely like a sinusoid that oscillates 1000 times per second. Notice the periodic peaks in the spectrum, denoting harmonic distortion of the original 1kHz pure tone. The spectrum of the original signal is a point spectrum at 1kHz.

This will be the standard example for the discussion on dithered quantization throughout this chapter. ♣

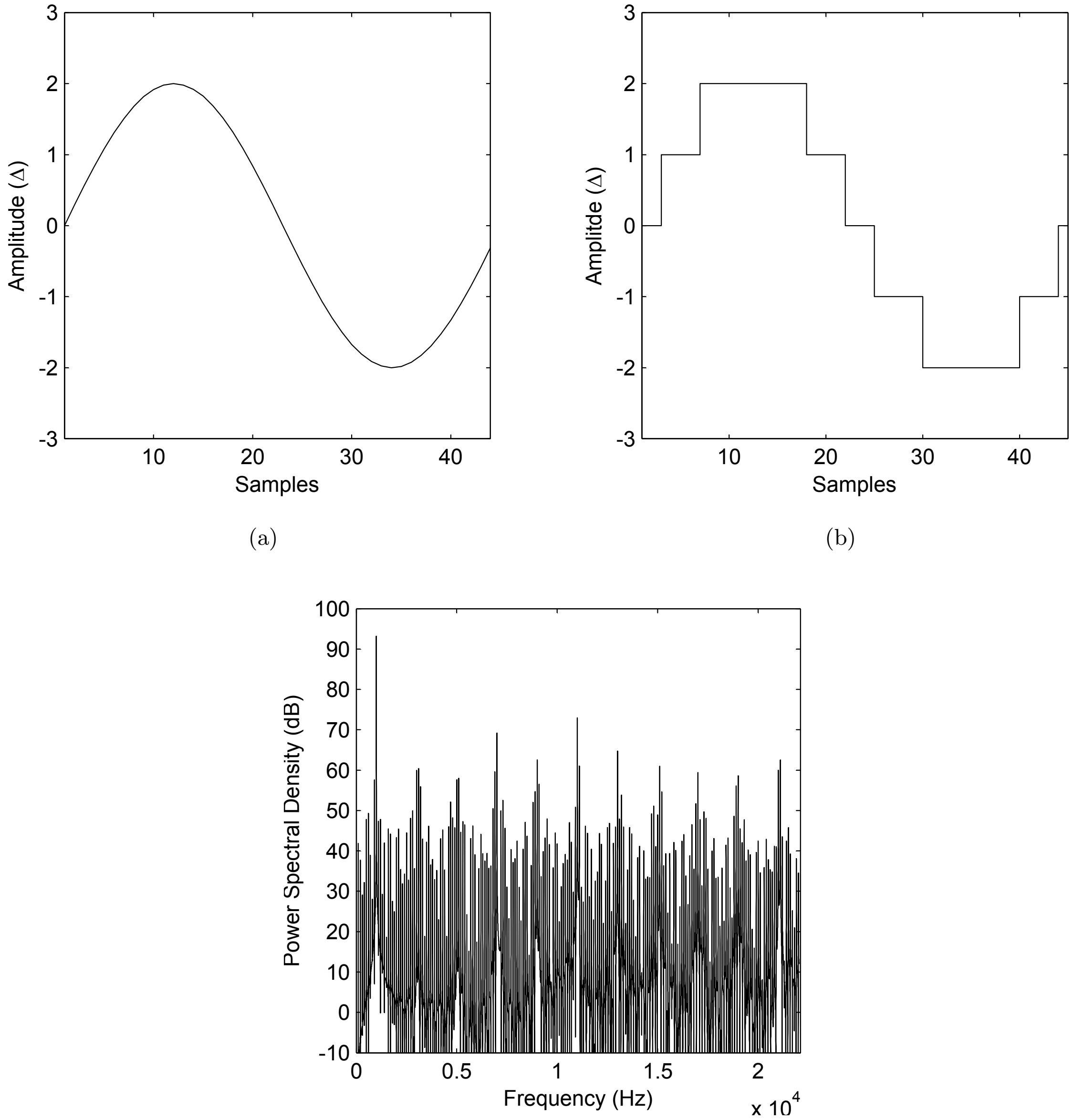
1.2 Subtractive Dither

Suppose we add an independent signal ν, called a dither signal, to our input signal x before quantizing. We will call this combined signal $w = x + \nu$. Then the output of the quantizing system is $Q(w)$. Thus, the quantization error of the system is

$$q = Q(x + \nu) - (x + \nu).$$

If we now subtract the independent signal ν from our system after quantization, then the output of this system becomes $y = Q(w) - \nu$. Thus, the total error of the dithered quantizer is equal to

$$\epsilon = Q(w) - \nu - x = q,$$



(a)

(b)

(c)

Figure 1.6: a) 1 period of the original signal. b) 1 period of the sampled and quantized signal. c) The spectrum of the quantized signal.

the quantization error. This procedure is shown schematically in Figure 1.7, and is called subtractively dithered quantization.

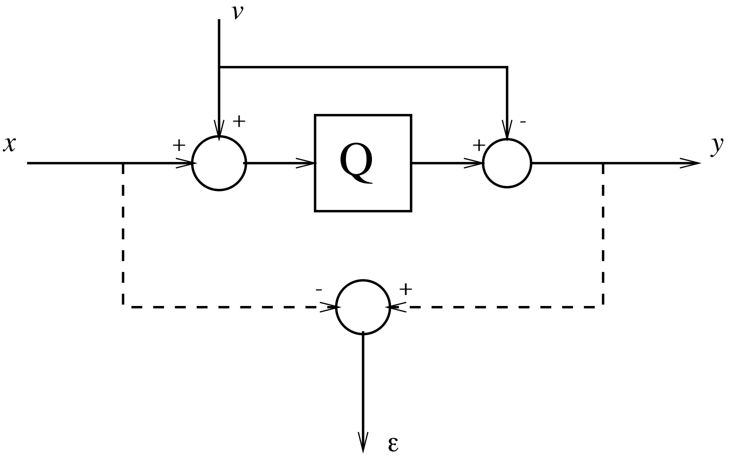


Figure 1.7: The subtractively dithered quantization schematic.

At this point, we are not free to choose the input signal x as we please, but we do have freedom with the choice of ν. Note that because the pdf of $x+\nu$, $p(x+\nu)$, is equivalent to the convolution of the the two pdfs, $p(x+\nu) = p(x) * p(\nu)$, assuming x and ν are independent, and the inverse Fourier transform of a convolution is a multiplication of the transforms of the individual pdfs, then the characteristic function of w is equivalent to the multiplication of the characteristic functions for x and ν. Hence, by choosing ν carefully such that $P_\nu(u) = 0$ for $u = \pm\frac{1}{\Delta}, \pm\frac{2}{\Delta}, \ldots$, we will have an input to an undithered quantizing system that satisfies the conditions of Theorem 1.1.5, and the moments of the total error of this system are given by Equation (1.1).

Theorem 1.2.1 (Schuchman's Condition [69]). *In a subtractively dithered quantizing system, the total error will be uniformly distributed and statistically independent of the input for arbitrary input distributions if and only if the characteristic function of the dither, P_ν, satisfies $P_\nu(u) = 0$ for $u = \pm\frac{1}{\Delta}, \pm\frac{2}{\Delta}, \ldots$.*

One such dither signal that satisfies the required conditions is a spectrally white dither with a uniformly distributed probability density function of width Δ. That is, the dither signal has a rectangular probability density function (RPDF). The pdf of this dither signal is shown in Figure 1.8. This is not the only dither signal which works, but this is the lowest power dither that satisfies Schuchman's Condition. It should be noted that this does not guarantee that the error is spectrally white. We need slightly more in order to satisfy the conditions of Theorem 1.1.6, which will tell us that the error is a white noise.

Theorem 1.2.2 ([72]). *In a subtractively dithered quantizing system, where ϵ_1 and ϵ_2 are two* *total error values separated in time by $\tau \neq 0$ with corresponding input values of x_1 and x_2, and dither values ν_1 and ν_2, respectively, the random vector (ϵ_1, ϵ_2) is statistically independent of the vector (x_1, x_2) for arbitrary input distributions if and only if*

$$P_{\nu_1,\nu_2}(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}) = 0$$

for $k_1 = 0, \pm1, \pm2, \ldots$, and $k_2 = 0, \pm1, \pm2, \ldots$, but not $(k_1, k_2) = (0, 0)$. Furthermore, if this condition holds, then

$$p_{\epsilon_1,\epsilon_2}(\epsilon_1, \epsilon_2) = \Pi_\Delta(\epsilon_1)\Pi_\Delta(\epsilon_2),$$

so that ϵ_1 and ϵ_2 are both uniformly distributed and independent of each other.

What this theorem tells us is that in order to guarantee that the total error of the quantized signal is spectrally white, each sample of our dither signal's values must be statistically independent [72]. In other words, the dither signal has to be a "random" signal with the correct pdf. Thus RPDF dither will render the total error of any input signal statistically independent and spectrally white – an independent, additive, white-noise process. Throughout the sequel, the term "dither" will implicitly mean a random signal.

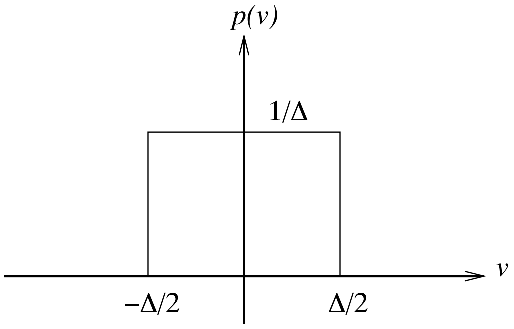
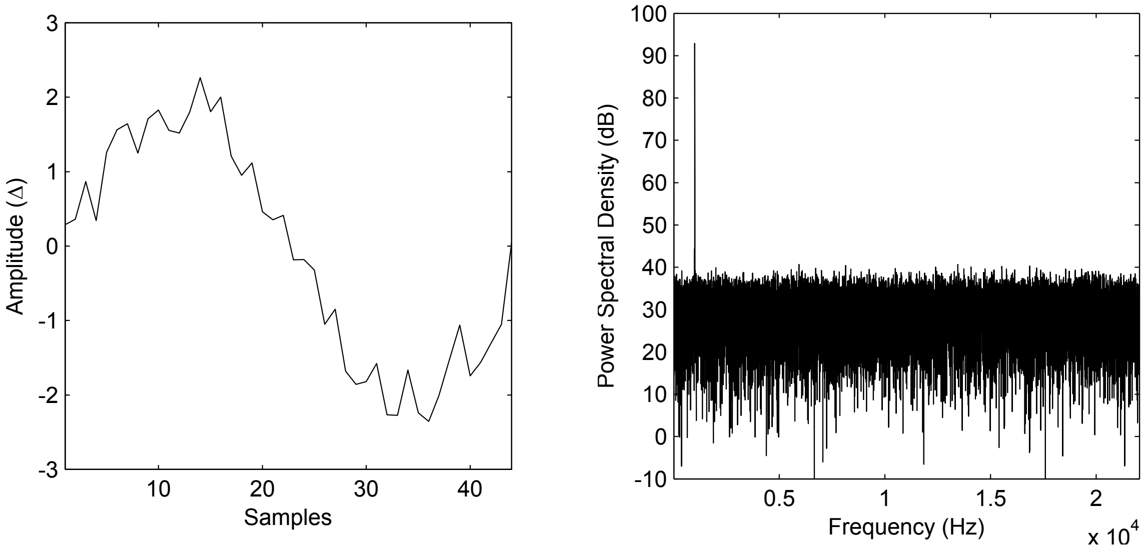


Figure 1.8: Rectangular PDF.

The most important limitation of subtractive dither is that it requires us to subtract the exact same dither signal after quantization. This effectively "de-quantizes" the quantized output, rendering it continuous in amplitude once again. In applications, this means that the quantization would be done from the side transmitting the data, and the subtraction would have to be done by the receiver. This requires one to synchronize the dither signal used across whatever media channel is used. Also, any processing of the quantized signal before the dither is subtracted will require us to perform similar operations on the dither signal. For these reasons, subtractive dither is not a viable option in practice. However, it is still of theoretical importance since it provides us with a benchmark with which to gauge other methods.

Example 1.2.3. We will turn back to our example of the 1kHz sine wave. For this example, we have taken our sine wave and applied the subtractive dither scheme using RPDF dither, shown in Figure 1.9(a). The spectrum of the output signal is shown in Figure 1.9(b). Note that the spectrum is white noise-like, with the exception of the single peak which represents the original 1kHz sine wave. The first moment of the error is $E[\epsilon^1] = 0$, and the second moment is $E[\epsilon^2] = \frac{\Delta^2}{12}$, as given by Equation (1.1).



(a)

(b)

Figure 1.9: a) 1 period of the subtractively dithered output signal. b) Spectrum of the subtractively dithered output signal

♣

1.3 Nonsubtractive Dither

The problem with subtractive dither was that one needed to keep track of the dither signal that was used in order to subtract it later. Since this is infeasible for most practical applications, alternative methods were studied. The first logical step is to study the properties of the quantization system if the dither signal is not subtracted after quantization. This is the essence of nonsubtractive dither. A schematic of this procedure is shown in Figure 1.10.

Unlike the subtractively dithered case, when using a nonsubtractively dithered quantization scheme, it is not possible to render the total error of the system statistically independent of the input. The best we can do is to render certain statistical moments of the total error independent of the input. The total error of the dithered quantizer is $\epsilon = Q(x + \nu) - x = q + \nu$.

Theorem 1.3.1 ([62, 73]). *In a nonsubtractively dithered quantization system, $E[\epsilon^m]$, the m^{th}*

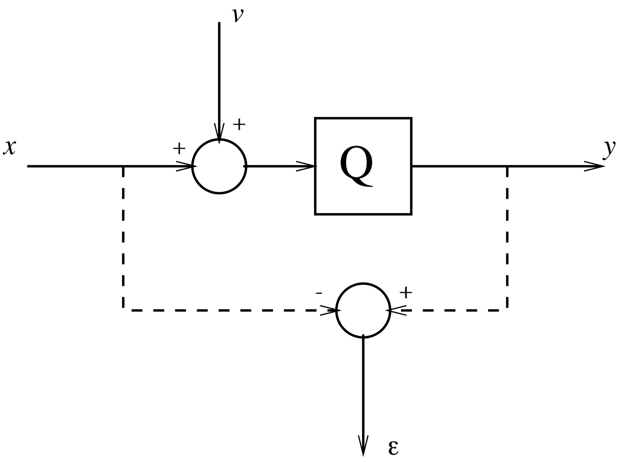


Figure 1.10: The nonsubtractively dithered quantization schematic.

moment of the error, is independent of the distribution of the system input if and only if

$$\frac{d^m}{du^m}\left(\frac{\sin(\pi\Delta u)}{\pi\Delta u}P_\nu(u)\right)\Bigg|_{u=\frac{k}{\Delta}} = 0$$

for $k = \pm 1, \pm 2, \ldots$.

If this theorem is satisfied for the index m, then the m^{th} statistical moment of the error is given by

$$E[\epsilon^m] = \left(\frac{i}{2\pi}\right)^m \frac{d^m}{du^m}\left(\frac{\sin(\pi\Delta u)}{\pi\Delta u}P_\nu(u)\right)\Bigg|_{u=0}. \tag{1.2}$$

If no dither is used, then P_ν is a constant, and the conditions of Theorem 1.3.1 will not be satisfied for any m. This agrees with our earlier comments on undithered quantization; that the total error of the system is not statistically independent of the input.

If we consider the RPDF dither signal that we used in the subtractively dithered case, then

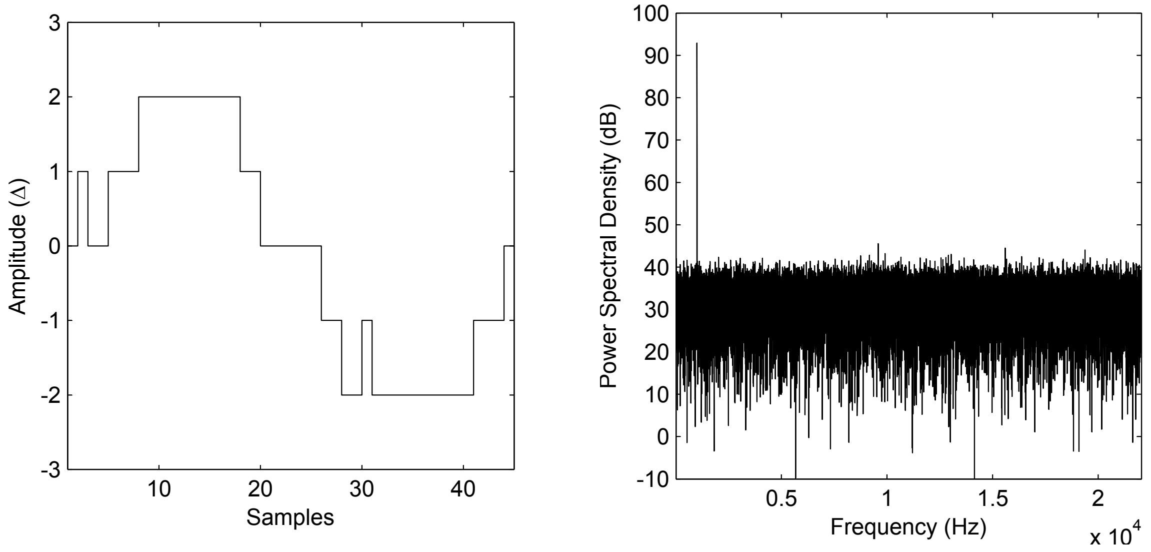
$$\frac{\sin(\pi\Delta u)}{\pi\Delta u}P_\nu = \left(\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right)^2.$$

The first derivative of this function is

$$2\frac{\sin(\pi\Delta u)}{\pi\Delta u}\left(\frac{\cos(\pi\Delta u)}{u} - \frac{\sin(\pi\Delta u)}{\pi\Delta u^2}\right),$$

which does evaluate to 0 when $u = \pm\frac{1}{\Delta}, \pm\frac{2}{\Delta}, \ldots$. This means that when using RPDF dither in a nonsubtractively dithered quantization system, the first moment of the total error is statistically independent of the input. In fact, the first moment of the error $E[\epsilon^1] = 0$. Notice that higher derivatives do not satisfy the conditions of Theorem 1.3.1, and so higher moments of the error are not controlled with this dither signal.

Example 1.3.2. Figure 1.11 shows the output of the nonsubtractive dither scheme when using our 1kHz sine wave and RPDF dither.



(a) (b)

Figure 1.11: a) 1 period of the nonsubtractively RPDF dithered output signal. b) Spectrum of the nonsubtractively RPDF dithered output signal.

♣

The fact that the second moment of the error is not controlled with RPDF dither is significant. By rendering the first statistical moment of the error independent of the input, we have succeeded in making the output sound like the input signal plus a white background noise. However, this noise is not benign. The noise will fluctuate in power along with the input signal. This fluctuation is determined by signal dependent variation of the second moment of the total error signal, and is commonly referred to as noise modulation in the literature.

In order to control the second moment of the total error signal, one uses a dither signal with a triangular probability density function (TPDF) with a width of 2Δ. The pdf of this dither signal is shown in Figure 1.12. TPDF dither satisfies the conditions of Theorem 1.3.1 for $m = 1$ and 2. At this point, we still have $E[\epsilon^1] = 0$, but now we also have $E[\epsilon^2] = \frac{\Delta^2}{4}$ (as derived from Equation (1.2)). Compare this with the case of subtractive dither. We expect an increase in the noise power over the subtractively dithered case.

For a detailed analysis and history of nonsubtractive dither, the reader is referred to works by Lipshitz, Vanderkooy, and Wannamaker [62, 73].

Example 1.3.3. This is our 1kHz sine wave again, this time nonsubtractively quantized after adding TPDF dither. Figure 1.13 shows the output of the system. Notice that the noise level

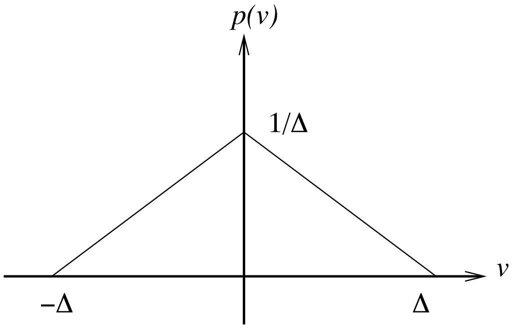
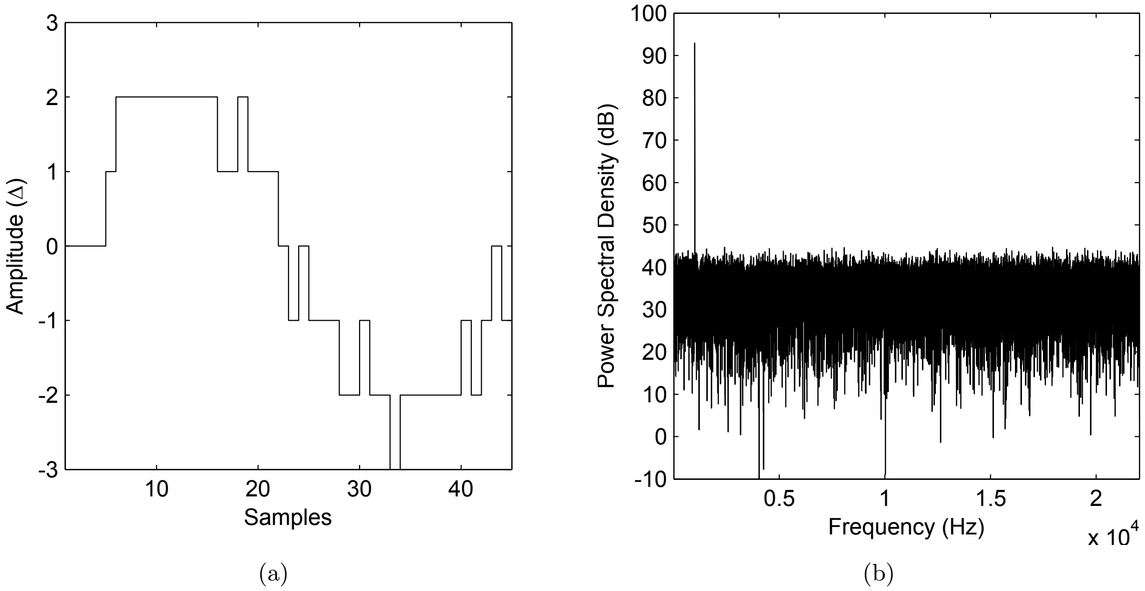


Figure 1.12: Triangular PDF.

is getting higher (the level of the white noise background is getting larger, as seen in Figure 1.13(b)) as we add larger dither signals.



(a)

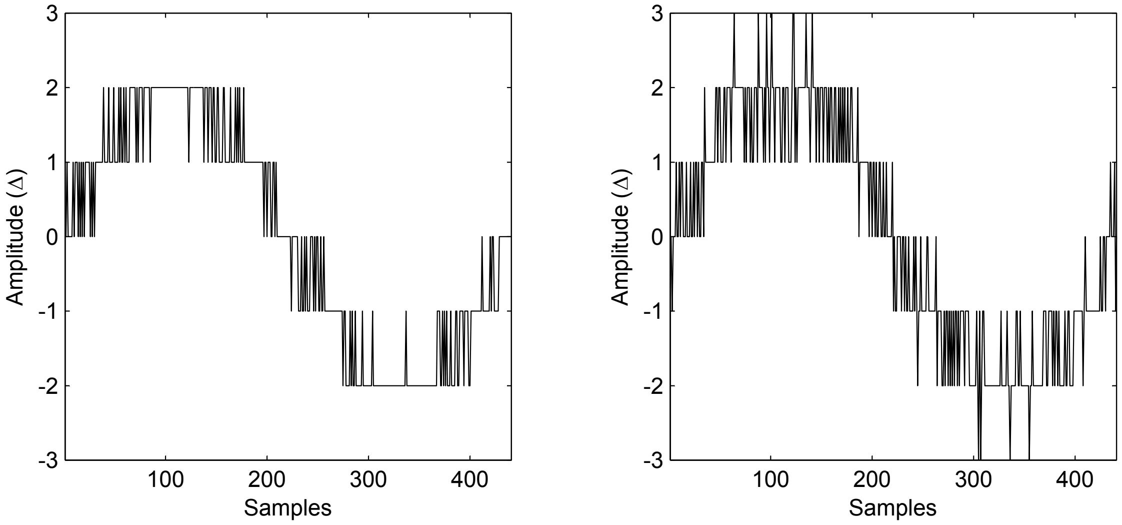
(b)

Figure 1.13: a) 1 period of the nonsubtractively TPDF dithered output signal. b) Spectrum of the nonsubtractively TPDF dithered output signal.

♣

Example 1.3.4. In order to show the difference between RPDF and TPDF a little more clearly, let us examine the case when we sample our 1kHz sine wave at 441kHz, an increase in the sampling rate by a factor of 10. Figure 1.14 shows the difference in output between the two dither schemes. Notice that the output of the RPDF dither scheme seems to jump between quantizer levels less frequently at the peaks of the sine wave, whereas the output of the TPDF dither scheme seems to jump around at a consistent rate. This is responsible for the varying loudness of the noise in the RPDF dither scheme, and the constant benign white noise in the TPDF dither scheme.

♣



(a)

(b)

Figure 1.14: a) 1 period of the output signal of the RPDF dither scheme. b) 1 period of the output signal of the TPDF dither scheme.

TPDF dither does not control the third or higher moments of the total error signal. However, after extensive testing on the subject, it is generally agreed that the human ear is not sensitive to statistical moments higher than the second. This means that trying to control total error moments higher than the second moment is unnecessary for audio applications, but could be relevant in some measurement applications.

1.4 Minimum-Phase Systems

Before we go further into the theory of dithered quantization, we will need to take a brief mathematical interlude. In this section, we will define some properties of 1-D sequences and systems. The notion of a minimum-phase system will also be discussed.

Definition 1.4.1. A one-sided sequence is a sequence, $x(n)$, such that $x(n) = 0$ for all $n < 0$. ⧫

A 1-D system, T, is a deterministic mapping defined on a suitable linear space of sequences . If T is a system and $x(n)$ is an input sequence, then $T(x(n))$ is well-defined. A linear system has the property that

$$T(a \cdot x_1(n) + x_2(n)) = a \cdot T(x_1(n)) + T(x_2(n)),$$

where a is a scalar. A shift-invariant system has the property that

$$T(x(n-m)) = y(n-m),$$

where $T(x(n)) = y(n)$. Roughly speaking, this means that the system is not sensitive to the position of the current value that it is operating on in the input sequence.

The systems that we will be concerned with are both linear and shift-invariant. This class of systems is nice because the behaviour of a linear shift-invariant system is entirely determined by its response to an impulse signal. We will often use $h(n) = T(\delta(n))$ to denote the impulse response of a system. For a summable sequence, $x(n)$, and a summable impulse response, we have

$$T(x(n)) = \sum_m x(m)T(\delta(n-m)) = x(n) * T(\delta(n)),$$

where $*$ denotes convolution and

$$\delta(n) = \begin{cases} 1, n = 0 \\ 0, \text{otherwise} \end{cases}$$

is the Kronecker delta. Unless explicitly stated otherwise, one can assume throughout the sequel that a system is linear and shift-invariant.

The inverse of a system, T, is a system which nullifies the effect of T. That is,

$$T(T^{-1}(x(n))) = T^{-1}(T(x(n))) = x(n).$$

In particular, if the impulse response of T^{-1}, $h^{-1}(n)$, is summable, then

$$h^{-1}(n) * h(n) = \delta(n).$$

If the impulse response of a system has finitely many non-zero entries, then it is referred to as a finite impulse response (FIR) system. Otherwise, it is referred to as an infinite impulse response (IIR) system.

Definition 1.4.2. A linear shift-invariant system is causal, or recursively computable, if its impulse response is a one-sided sequence. ⧫

A system is said to be stable if its output doesn't escape to arbitrarily large numbers when given reasonable input values. For this application, "reasonable" will mean bounded input signals.

Definition 1.4.3. A linear shift-invariant system is stable if and only if its impulse response is absolutely summable. That is

$$\sum_n |h(n)| < \infty. \qquad \blacklozenge$$

This is referred to as bounded input-bounded output (BIBO) stability. For systems with a large, or infinite, impulse response, it can be infeasible to compute this sum directly in order to check for stability. For a more efficient way of determining stability, we turn to transfer functions of a system [61].

Definition 1.4.4. The z-transform of a sequence $x(n)$ is defined as

$$X(z) = \mathscr{Z}[x(n)](z) = \sum_n x(n)z^{-n}. \qquad \blacklozenge$$

Theorem 1.4.5 (Convolution Theorem). *Given two summable sequences, $x_1(n)$ and $x_2(n)$, the z-transform of their convolution is given by*

$$\mathscr{Z}[x_1(n) * x_2(n)](z) = X_1(z)X_2(z).$$

The transfer function of a system is given by the ratio of the transformed output sequence over the transformed input sequence [59]. For discrete systems (systems that act on sequences instead of continuous functions), we use the z-transform. For a linear shift-invariant system, recall that

$$y(n) = x(n) * h(n).$$

Taking the z-transform of both sides yields

$$Y(z) = X(z)H(z) \Rightarrow \frac{Y(z)}{X(z)} = H(z).$$

So the transfer function of a linear shift-invariant system is the z-transform of its impulse response.

Theorem 1.4.6. *A linear shift-invariant system with the transfer function $H(z)$ is stable if and only if its poles are located inside the unit circle $|z| = 1$.*

This theorem provides a much easier test than trying to compute the sum of the magnitude of the impulse response of a system. If $H(z)$ has a pole located on the unit circle (but does not have any located outside the unit circle), then this system is referred to as marginally stable. This is a limiting case between stable and unstable systems.

Given a system, one often wants to know about the stability of its inverse.

Theorem 1.4.7. *The inverse of a linear shift-invariant system with transfer function $H(z)$ is stable if and only if all the zeros of $H(z)$ are located inside the unit circle.*

Proof. The transfer function of the inverse system is $\frac{1}{H(z)}$. Therefore, the zeros of $H(z)$ are the poles of $\frac{1}{H(z)}$; hence the inverse system is stable if and only if all the zeros of $H(z)$ are located inside the unit circle. ■

Definition 1.4.8. A minimum-phase system is a linear shift-invariant system that is causal and stable with a causal and stable inverse. ♦

In 1-D systems, there are many equivalent statements which are used as alternative definitions of a minimum-phase system. Among the more popular equivalent definitions is the following:

Theorem 1.4.9. *A system is minimum-phase if it is causal, and its transfer function has all of its poles and zeros inside the unit circle,* $|z| = 1$.

Proof. This follows immediately from the definition of minimum-phase and Theorem 1.4.7. ■

Note that there is also the case of marginally minimum-phase systems corresponding, like the marginally stable case, to having zeros located on the unit circle (but none lying outside).

We have chosen our definition of minimum-phase because it is the most meaningful when extended to higher dimensions. We will see that talking about the locations of the poles and zeros of multidimensional systems becomes very complicated, but we can still talk about the stability of a system and its inverse.

In addition to identifying whether a system is stable or not, the transfer function of a system also provides us with additional information. Given a system with transfer function $H(z)$, the graph of $\log(|H(e^{i\theta})|)$, called the log-magnitude response of the system, tells us how the system will amplify a sinusoidal signal of frequency $\frac{\theta f_s}{2\pi}$. Also, $arg(H(e^{i\theta}))$ tells us the phase lead that the system will add to a sinusoidal input. A system will usually be identified by its impulse response, its transfer function, or by its log-magnitude and phase.

1.5 Noise Shaping and Jensen's Theorem

So far, we have been able to show how to create a quantized signal which behaves qualitatively like the original signal with an additive white noise. However, is this additive white noise an optimal choice? What are the perceptual qualities of error signals with other spectra? To change the spectrum of the error signal, we employ the following error feedback circuit, called a noise shaper [74]. The noise shaping schematic is shown in Figure 1.15.

The filter H, which is assumed to be stable, is defined to delay the signal by at least one time step. This circuit takes the total error of the dithered quantizer, ϵ, and subtracts this from the subsequent values of the input sequence $x(n)$, becoming the sequence $x'(n) = x(n) - h(n) * \epsilon(n)$.

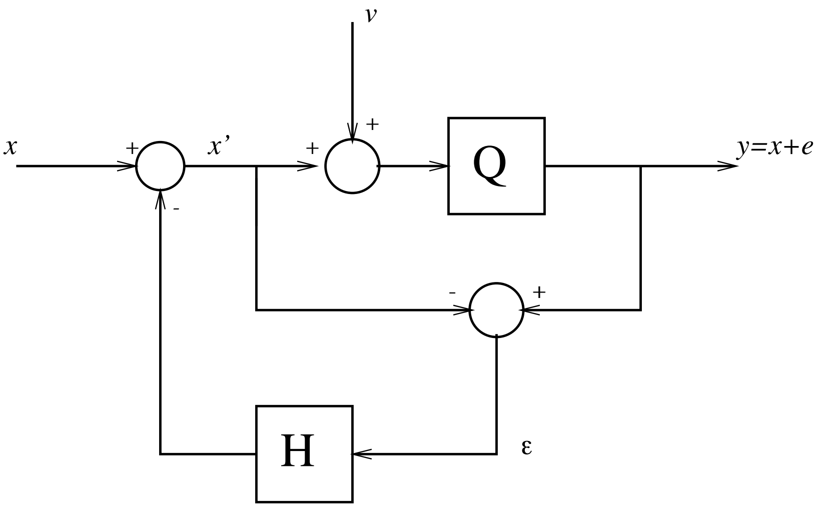


Figure 1.15: The noise shaping schematic.

The signal $x'(n)$ is then fed into the dithered quantizer, and so the output of the system is $x'(n) + \epsilon(n) = x(n) + e(n)$, where $e(n)$ denotes the total error of the system. We have

$$x'(n) = x(n) + e(n) - \epsilon(n), \tag{1.3}$$

and

$$x'(n) = x(n) - h(n) * \epsilon(n). \tag{1.4}$$

Subtracting Equation (1.4) from Equation (1.3) gives

$$e(n) = \epsilon(n) - h(n) * \epsilon(n).$$

Taking the z-transform gives

$$\Xi(z) - H(z)\Xi(z) = E(z) \Rightarrow \frac{E(z)}{\Xi(z)} = 1 - H(z).$$

Hence, the quantity $1 - H(z)$ is referred to as the noise transfer function. It relates the total error of the dithered quantizer, ϵ, to the total error, e, of the output of the noise shaper. In accordance with the theory of dithered quantization presented earlier, $\epsilon(n)$ has a white spectrum with zero mean that does not fluctuate in power as long as we dither the system correctly (i.e. TPDF dither). Therefore, the spectrum of the total error of this system is determined by the noise transfer function, $1 - H(z)$. If the system is not properly dithered, then signal dependent distortions can occur in the error signal that are undesirably audible, such as "chirps" caused by limit-cycle oscillations in the feedback loop, and noise modulation.

Example 1.5.1. Consider our 1kHz sine wave. This time, we will use a simple single delay noise shaping system where $H(z) = z^{-1}$. Therefore the noise transfer function is $1 - H(z) = 1 - z^{-1}$.

This particular system takes the the total quantizer error and subtracts this from the next input value (a single sample time delay). This filter, $H(z)$, would not actually be used in audio applications since better systems can be designed with little added effort. However, this will suit to illustrate the mechanics of the noise shaping technique.

The log-magnitude response of the system $1 - H(z)$ is shown in Figure 1.16. The output of the noise shaping system is shown in Figure 1.17. Note that the noise of the output spectrum is determined by the log-magnitude response of $1 - H(z)$. This is true because the system is properly dithered. If we do not use TPDF dither, undesirable artifacts can appear in the output. Figure 1.18 shows the spectrum of the output when no dither is used.

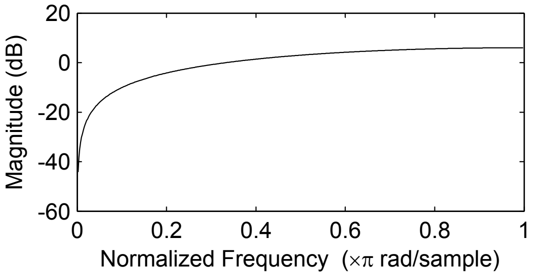
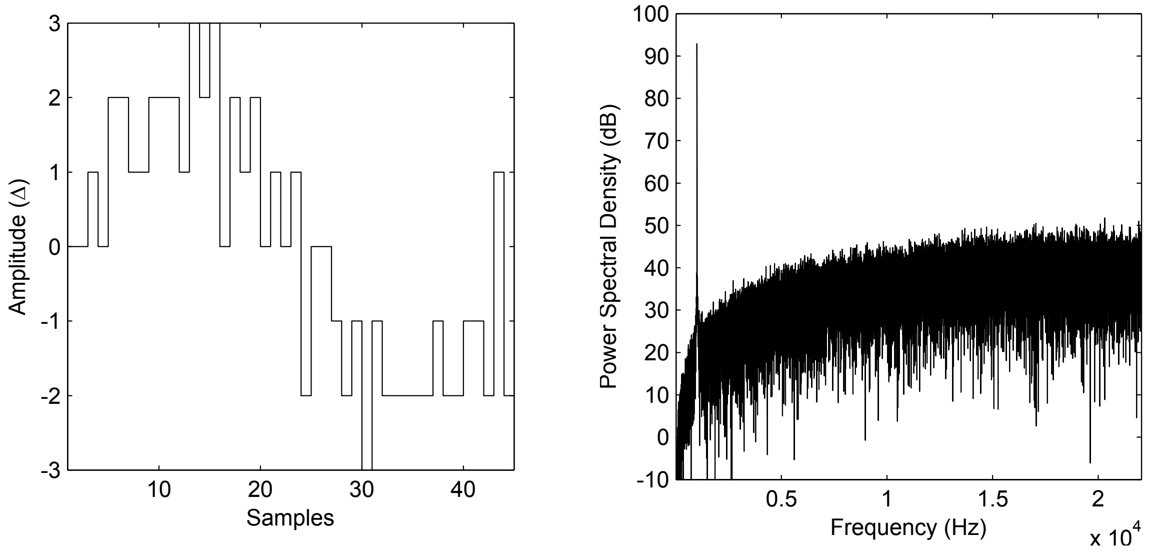


Figure 1.16: Log-magnitude of $1 - H(z)$ from Example 1.5.1.



(a)

(b)

Figure 1.17: a) 1 period of the noise shaped output signal. b) Spectrum of the noise shaped output signal.

♣

Now that we know how to alter the spectrum of the total error almost as we please, we

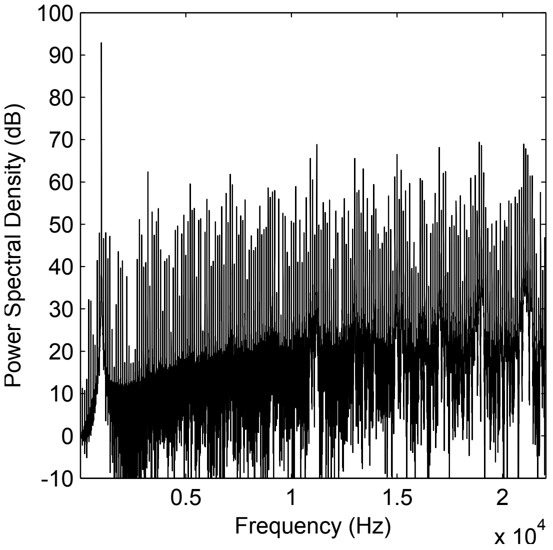


Figure 1.18: Spectrum of noise shaped output signal if no dither is used.

would like to use this knowledge to reduce the audibility of the error signal. Figure 1.19 shows the sensitivity of the human ear to low-level noise as a function of the frequency of the sound [82]. The higher the sensitivity, the louder that sound will seem to the human listener. Note that the ear has a maximal sensitivity around 3kHz. This is the resonance frequency of the ear canal. Also, the ear's sensitivity drops very quickly for higher frequency sounds. For acoustic applications, we wish to invert this curve and design a system of the form $1 - H(z)$ that has the inverted shape. We will come back to this point in the next section.

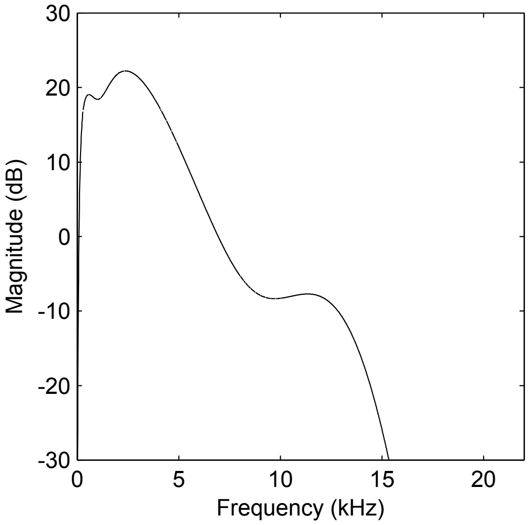


Figure 1.19: The sensitivity of the human ear as a function of frequency.

There are several different systems with a transfer function of the form $1 - H(z)$ that can produce the same log-magnitude shape.

Example 1.5.2. Consider the following transfer functions:

$$1 - 6z^{-1} + 8z^{-2}, \tag{1.5}$$

$$1 - 4.5z^{-1} + 2z^{-2}, \tag{1.6}$$

$$1 - 2.25z^{-1} + 0.5z^{-2}, \tag{1.7}$$

$$1 - 0.75z^{-1} + 0.125z^{-2}. \tag{1.8}$$

The log-magnitude of these transfer functions can be seen in Figure 1.20. Note that any value above 0dB corresponds to an amplification of noise volume in our noise shaping circuit, and values below 0dB correspond to a suppression of noise volume. The top log-magnitude in the figure is the log-magnitude associated with Equation (1.5), the next highest log-magnitude corresponds to Equation (1.6), and so on respectively. All log-magnitudes have the same shape, but they have differing amounts of amplification and suppression. We will discover shortly that the amount of suppression possible for a given log-magnitude shape depends on the location of the zeros of the transfer function.

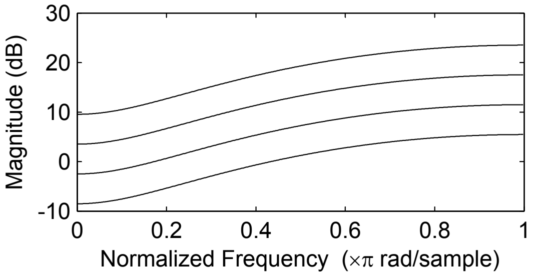


Figure 1.20: Several transfer functions with the same log-magnitude shape.

♣

So designing a system, $1 - H(z)$, with the characteristics we desire does not have a unique solution. However, there is a unique system whose log-magnitude shape will be as desired, but which will also suppress as much noise as possible.

Theorem 1.5.3 (Jensen's Inequality [33]). *Given a causal stable FIR system with transfer function $1 - H(z)$, where $H(z)$ has no constant term, we have*

$$\int_{-\pi}^{\pi} \log\left(\left|1 - H(e^{i\theta})\right|\right) d\theta \geq 0,$$

with equality if and only if $1 - H(z)$ is the transfer function of a minimum-phase system.

This theorem tells us that the minimum-phase system suppresses the largest amount of noise out of the possible system choices with the same log-magnitude shape. This theorem is actually a corollary of the much more powerful Jensen's Theorem, which gives an explicit relationship between the location of the zeros of a transfer function and the area under the log-magnitude graph.

Theorem 1.5.4 (Jensen's Theorem [33]). *Given a causal stable FIR system with transfer function $1 - H(z)$, where $H(z)$ has no constant term, we have*

$$\int_{-\pi}^{\pi} \log\left(\left|1 - H(e^{i\theta})\right|\right) d\theta = 2\pi \sum_j \log\left(|\alpha_j|\right),$$

where α_j are all zeros of $1 - H(z)$ located outside the circle $|z| = 1$.

Example 1.5.5. Recall Example 1.5.2. Equation (1.5) has zeros at $z = 2$ and $z = 4$. The remaining transfer functions were formed by reflecting these zeros inside the unit circle to $z = \frac{1}{2}$ and $z = \frac{1}{4}$. Equation (1.8) is the only transfer function with all of its zeros inside the unit circle. In accordance with Jensen's Theorem, this transfer function also offers the greatest possible amount of noise suppression. Also notice that this curve has equal areas above and below 0dB, and that the other curves obey Jensen's Inequality. ♣

Note that if we have a transfer function of the form $a - H(z)$, for a non-zero constant a, then

$$\int_{-\pi}^{\pi} \log\left(\left|a - H(e^{i\theta})\right|\right) d\theta = \int_{-\pi}^{\pi} \log(|a|) + \log\left(\left|1 - \frac{H(e^{i\theta})}{a}\right|\right) d\theta \geq 2\pi \log(|a|)$$

with equality if and only if $1 - \frac{H(z)}{a}$ is minimum-phase. However, multiplying a transfer function by a non-zero constant does not change its stability, or the stability of its inverse. Therefore, we can say that equality is attained if and only if $a - H(z)$ is minimum-phase.

Before we move on, consider the case where $a - H(z)$ contains complex coefficients. Jensen's theorem still holds under these conditions. These systems are not of practical importance. However, they are of theoretical use and will feature in our theory later on. If all the coefficients of $a - H(z)$ are real, then the log-magnitude of this system is an even function. Hence, one might see Jensen's inequality written as [81]

$$\int_{0}^{\pi} \log\left(\left|1 - H(e^{i\theta})\right|\right) d\theta \geq 0$$

If we allow for complex coefficients, then the log-magnitude of the system is no longer even, and we need to integrate over the entire interval, $[-\pi, \pi]$, for Jensen's Theorem to hold true.

1.6 The Hilbert Transform Method for Designing Noise Shapers

There are a number of available methods to design a minimum-phase system with a transfer function of the form $1-H(z)$ whose log-magnitude has been specified. The method we are going to focus on is the Hilbert transform. In this section, the Hilbert transform will be defined, and we will show that the log-magnitude and phase of a minimum-phase system are related by the Hilbert transform. Afterwards, we will show how one numerically calculates the minimum-phase phase that corresponds to a given log-magnitude.

Consider a piecewise continuous function, $f(t)$. This function can be uniquely decomposed into even and odd components:

$$\begin{aligned} f_e(t) &= \frac{f(t)+f(-t)}{2}, \\ f_o(t) &= \frac{f(t)-f(-t)}{2}. \end{aligned}$$

Thus, $f(t) = f_e(t) + f_o(t)$, and $f_e(t)$ and $f_o(t)$ are even and odd functions, respectively.

If the function $f(t)$ is such that $f(t) = 0$ for $t < 0$, then $f(t)$ is referred to as a causal function. Causal functions are of interest because in this case we can write an explicit relationship between $f_e(t)$ and $f_o(t)$. Specifically

$$f_o(t) = sgn(t) f_e(t),$$

where

$$sgn(t) = \begin{cases} 0, & t = 0 \\ 1, & t > 0 \\ -1, & t < 0 \end{cases}.$$

In particular, this allows us to write

$$f(t) = (1 + sgn(t)) f_e(t).$$

Definition 1.6.1. The Fourier transform of an integrable function, $f(t)$, is defined to be [60]

$$\mathcal{F}\{f(t)\} = F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi t\nu} dt.$$

The inverse Fourier transform is

$$\mathcal{F}^{-1}\{F(\nu)\} = f(t) = \int_{-\infty}^{\infty} f(t) e^{i2\pi t\nu} d\nu.$$

♦

Now, consider the Fourier transform of the causal function $f(t)$.

$$\mathcal{F}\{f(t)\} = \mathcal{F}\{f_e(t)\} + \mathcal{F}\{sgn(t)\} * \mathcal{F}\{f_e(t)\},$$

where the convolution arises from the convolution theorem for the Fourier transform. The Fourier transform of $sgn(t)$ is known to be

$$\mathcal{F}\{sgn(t)\} = \frac{-i}{\pi\nu}.$$

Therefore, for sufficiently nice functions f, we have

$$\mathcal{F}\{f(t)\} = F(\nu) = F_e(\nu) - i\text{P.V.}\int_{-\infty}^{\infty} \frac{F_e(s)}{\pi(\nu - s)} ds,$$

where P.V. stands for the principle value of this integral.

Definition 1.6.2. The convolution $F(\nu) * \frac{-1}{\pi\nu}$ is defined to be the Hilbert transform of the function $F(\nu)$. The inverse Hilbert transform is defined to be the a convolution with $\frac{1}{\pi\nu}$. ♦

For a real-valued even function, $f_e(t)$, the Fourier transform, $\mathcal{F}\{f_e(t)\} = F_e(\nu)$, is real. Therefore, following the definition of the Hilbert transform, we can say that the real and imaginary parts of the Fourier transform of a causal function are related by the Hilbert transform. This statement actually works in reverse as well. That is, if the real and imaginary parts of a complex valued function, $F(\nu)$, $\nu \in \mathbb{R}$, are related by the Hilbert transform, then the inverse Fourier transform of $F(\nu)$ is a causal function [27].

Another important property of the Hilbert transform is that if a complex function, $F(s)$, is analytic in the right half of the s-plane, $\mathcal{R}(s) > 0$, and periodic of period 2π along the imaginary axis, then the real and imaginary parts of this function along the imaginary axis are also related by the Hilbert transform. Using the conformal mapping

$$z = e^s$$

we can also deduce that if a complex function, $F(z)$, is analytic in the region outside the unit circle in the z-plane, $|z| > 1$, then the real and imaginary parts of $F(e^{i\theta})$ are also related by the Hilbert transform.

Now that we have introduced the Hilbert transform, the next step is to demonstrate that the log-magnitude and phase of a minimum-phase system are related by the Hilbert transform.

Recall the z-transform, and let $X(z)$ be the z-transform of a causal sequence, $x(n)$. Note that $X(z)$ is, in general, an infinite sum, and we have to ask for which values of z does this sum converge. For any causal sequence, the region of convergence of $X(z)$ is the region $|z| > R$, for some R. The region may or may not include the circle $|z| = R$. Since $X(z)$ is a power series in z^{-1}, this function is analytic wherever the sum converges. Hence, $X(z)$ is analytic in the region $|z| > R$ whenever $x(n)$ is causal.

Now suppose $x(n)$ is the impulse response of a causal system. Then $X(z)$ is the transfer function of the system. If the region of convergence of $X(z)$ contains the unit circle, then the system is stable. In particular, a causal and stable system has a transfer function which is analytic for all $|z| > R$, for some $0 \leq R < 1$.

Note that the singularities of the inverse system include the zeros of the original system. Therefore, if a system is minimum-phase, then the transfer function $X(z)$ has no poles or zeros outside the unit circle. Hence, the logarithm of the transfer function is analytic in the region $|z| > R$, for some $0 \leq R < 1$. Recall that for complex valued functions, $\log(X(z)) = \log(|X(z)|) + i \arg(X(z))$. In particular, this means that $\log(|X(e^{i\theta})|)$ and $\arg(X(e^{i\theta}))$ are related by the Hilbert transform. These two quantities are precisely the log-magnitude and phase of the minimum-phase system.

Now that we know that the log-magnitude and phase of a minimum-phase system are related by the Hilbert transform, we can use this to find the impulse response of a system with a given log-magnitude.

In order to numerically compute the minimum-phase phase associated with a given log-magnitude, we use the discrete Hilbert transform. The definition of the discrete Hilbert transform is based on the principle that the Fourier transform of the even and odd parts of a causal function are related by the Hilbert transform.

Definition 1.6.3. The discrete Fourier transform of a sequence of length N, $x(n)$ is defined to be

$$\mathcal{F}\{x(n)\} = X(k) = \sum_{n=0}^{N-1} x(n) e^{-i2\pi \frac{nk}{N}}.$$

The inverse discrete Fourier transform is

$$\mathcal{F}^{-1}\{X(k)\} = x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{i2\pi \frac{nk}{N}}.$$

♦

Definition 1.6.4. A finite length sequence, $x(n)$, of length N, N even, is causal if $x(n) = 0$ for $\frac{N}{2} < n < N$. ♦

We can decompose a finite length sequence into even and odd parts just as we did with causal functions.

$$x_e(n) = \frac{x(n) + x(N-n)}{2}, n = 0, \ldots, N-1$$
$$x_o(n) = \frac{x(n) - x(N-n)}{2}, n = 0, \ldots, N-1$$

where $x(N)$ is defined to be equal to $x(0)$. Notice that $x_e(n) = x_e(N-n)$, and $x_o(n) = -x_o(N-n)$, so these sequences exhibit the similar notions of even- and odd-ness. In fact, the discrete Fourier transform of an even finite length sequence is real, just as the Fourier transform of an even function is real.

From this point, we will assume that N is an even integer. This is largely for convenience with the Fast Fourier Transform methods in MATLAB. For a causal finite length sequence, we deduce that

$$x_o(n) = sgn(n)x_e(n),$$

where

$$sgn(n) = \begin{cases} 1, & n = 1, \ldots, \frac{N}{2} - 1 \\ 0, & n = 0, \frac{N}{2} \\ -1, & n = \frac{N}{2} + 1, \ldots, N-1 \end{cases}.$$

The definition of the sgn function will change slightly if we allow N to be odd, but all of the results are still valid.

Motivated by the idea that the real and imaginary parts of the Fourier transform of a causal function are related by the Hilbert transform, we define the real and imaginary parts of the discrete Fourier transform of a causal finite length sequence to be related by the discrete Hilbert transform.

Definition 1.6.5. The discrete Hilbert transform is equivalent to a convolution with the discrete Fourier transform of $sgn(n)$. The specific form of this Fourier transform depends on whether the length, N, of the sequence is odd or even. ♦

In practice, we will use the causality relation of the discrete Hilbert transform instead of trying to do the actual convolution. That is, given a sequence, $x(n)$, we find the sequence $x(n) + iy(n)$, where $y(n)$ is the Hilbert transform of $x(n)$, by performing the inverse discrete Fourier transform on $x(n)$, multiplying this sequence by $1 + sgn(n)$, and then performing the discrete Fourier transform. The sequence $1 + sgn(n)$ is called the causalizer since it turns a given sequence, $s(n)$, into a causal sequence, $s(n)(1 + sgn(n))$, by pointwise multiplication.

Suppose that the sequence $X(k)$ is a set of sampled points from one period of the periodic function $\log(|H(e^{i\theta})|)$. Then the discrete Hilbert transform is the equivalent of evaluating the convolution integral in Definition 1.6.2 for periodic functions by the trapezoid rule, and produces an approximation to the minimum-phase phase associated with the continuous function

$\log(|H(e^{i\theta})|)$. This approximation converges to the actual minimum-phase phase, $arg(H(e^{i\theta}))$, in the limit as the number of samples N approaches ∞. The rate of pointwise convergence is approximately $O(\frac{1}{N^2})$ for almost any given log-magnitude (i.e. has continuous second derivatives at all but a finite number of points)[78].

In order to find the impulse response of a minimum-phase system given its sampled log-magnitude, $X(k)$, we follow these steps:

1. Compute the inverse discrete Fourier transform of $X(k)$, $\mathcal{F}^{-1}\{X(k)\} = x(n)$.
2. Causalize the resulting sequence $x(n)$; $x_c(n) = x(n) \cdot (1 + sgn(n))$.
3. Compute the discrete Fourier transform of $x_c(n)$. We now have $X(k)$ and an approximation to the sampled minimum-phase phase. In other words, we have a sampled form of $\log(|H(e^{i\theta})|) + i \arg(H(e^{i\theta})) = \log(H(e^{i\theta}))$.
4. Exponentiate each term in the sequence. We now have a sampled version of $H(e^{i\theta})$.
5. Compute the inverse discrete Fourier transform to get an approximation to the causal impulse response of the minimum-phase system.

Written out in mathematical shorthand, we can obtain the impulse response of the minimum-phase system by computing

$$h(n) = \mathcal{H}\{X(k)\}(n) = \mathcal{F}^{-1}\left\{\exp\left(\mathcal{F}\left\{\mathcal{F}^{-1}\{X(k)\} \cdot (1 + sgn(n))\right\}\right)\right\}(n). \tag{1.9}$$

This procedure will converge to the impulse response of a minimum-phase system with the transfer function $e^a - H'(z)$, where a is the average value of the log-magnitude, at a rate of $O(\frac{1}{N^2})$. That is, each component of the impulse response will converge at a rate of $O(\frac{1}{N^2})$. In order to obtain a system with a transfer function of the form $1 - H(z)$, as is required for our noise shaping theory, simply divide through by e^a after applying this Hilbert transform procedure, or ensure that the average value of the log-magnitude is zero before applying the procedure; either method will give the same result.

Example 1.6.6. To end this section, we will apply this Hilbert transform method to design a noise shaper for audio applications. Recall Figure 1.19. In order to create a noise shaper that suppresses noise where it is most audible to the human ear, we invert this graph and use the resulting shape as our target log-magnitude. However, the sharp drop at high frequencies causes an unnecessarily long impulse response. The actual target log-magnitude that we use is given in Figure 1.21. This is an IEEE standard for the sensitivity of the human ear.

The resulting impulse response is quite long. If we window the impluse response down to only 10 non-zero values using a Hann window (so that the log magnitude of the realized system does not contain ripples from truncation of the impulse response) [82], we obtain the transfer function

$$1-2.412z^{-1}+3.370z^{-2}-3.937z^{-3}+4.174z^{-4}-3.353z^{-5}+2.205z^{-6}-1.281z^{-7}+0.569z^{-8}-0.0847z^{-9}.$$

The log-magnitude of this system is given in Figure 1.22. This displays reasonably good agreement with our target design. If we were willing to create a noise-shaper of higher order (allow for greater powers of z^{-1}, we could match the target design shape as closely as we like (pointwise). Notice that the areas above and below the 0dB line of our noise shaper's log-magnitude are equal, in agreement with Jensen's Theorem. ♣

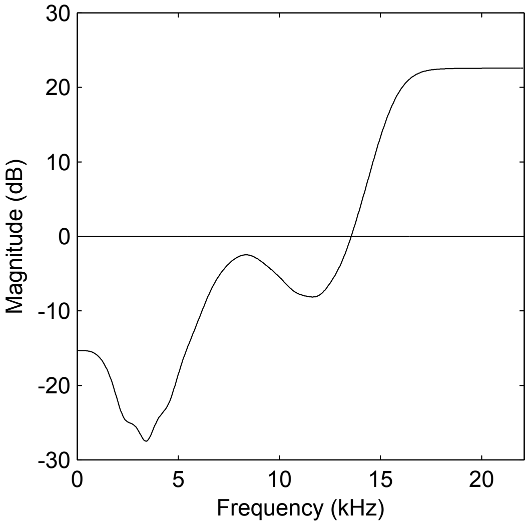


Figure 1.21: Target shape for the log-magnitude of an audio noise shaper.

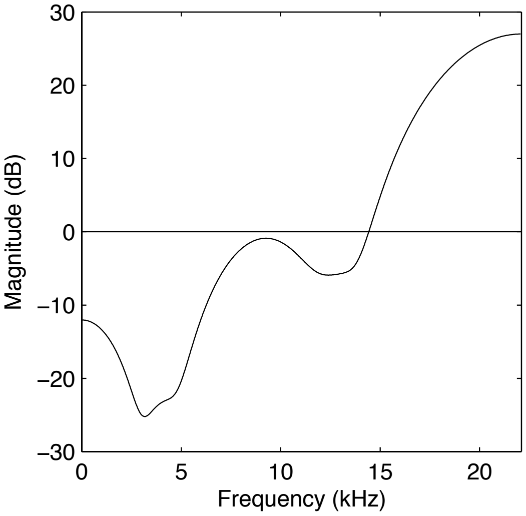


Figure 1.22: Log-magnitude of the designed noise shaper.

Chapter 2

Dither and Quantization of Images

Dithered quantization of images can be traced back to Schroeder [67] in 1969. In this paper, Schroeder introduces RPDF dithered quantization of images and presents a short discussion about the qualitative results. A similar study was done by Lippel and Kurland [47] in 1971, however they do not specify what type of random dither they are adding to the signal. At this point, the image processing community seems to have focussed their attention on the half-toning problem; that is quantization to only 1-bit. Dithered quantization, as we have developed it in this thesis, does not apply to the half-toning problem since we cannot dither the signal properly without saturating the quantizer, thereby destroying the statistical benefits that RPDF and TPDF dither would otherwise provide [34]. In 1975, Floyd and Steinberg [24] introduced noise shaping to image processing. They called the technique “error diffusion”. In 1988, Ulichney [80] briefly comments on dithered quantization in the context of the half-toning problem. The mention is only brief since, as mentioned before, proper dithering in our sense does not apply to 1-bit quantization. He illustrates other techniques that offer much better results for 1-bit quantization. Curiously, Ulichney does not mention what type of random dither is used in his examples. Once it was understood how the Floyd and Steinberg noise shaper worked on the spatial frequencies of an image, there were attempts to try to improve upon their design [45]. Throughout the literature are mentions of dithered quantizations in the sense that a random dither signal is added before quantization [40]. Often, random dither is introduced as an attempt to break up texture patterns and other undesirable artifacts. Rarely does this dither signal have the proper pdf as outlined in the previous chapter. Furthermore, the only author we have found who discusses the moments of the error signal in images is Brinton [11].

The focus of this section will be to demonstrate the effect of quantization and dither on images. In our series of demonstrations, the input signal will be a 8-bit monochrome image. That is to say that the brightness level of each sample point can have values between 0 and 255 in terms of least significant bits of the original image. 0 corresponds to black, and 255 corresponds to white. Figure 2.4 shows the actual image that will be used. The original image is 585×585 pixels (keep in mind that this is the dimension of the input signal and has nothing to do with the size of the transform used for the discrete Hilbert transform, as discussed in Section 1.6). All images are shown at the end of this chapter.

For these demonstrations, we have quantized the image to 3-bits. Hence, we are trying to recreate the original image as accurately as possible while only using 8 different shades of grey. The quantizer used for this demonstration is a "midriser" type, where the middle brightness value of 127.5 is assumed to be an input of 0. That is, the quantizer is centred around the middle brightness level.

The image used for this demonstration does not go to very dark or very light shades. The reason for this is that our theory of quantization, dither, and noise-shaping is only valid if we do not saturate the quantizer. Since we cannot display brightness values lower than 0 or higher than 255, we have to ensure that our output does not go beyond these thresholds. This means that we have to shrink the brightness range of the original image to allow enough headroom, both above and below the image, for our quantization and dither theory to work. The amount of headroom required depends on the amount of noise suppression one wishes to achieve with noise shaping; more drastic noise shaper designs require more headroom. For the noise shapers we will employ, we use about 2Δ of headroom around our input signal.

This may seem like a severe restriction to our theory, but quantizing an image to 3-bits is a very coarse quantization for the requantized image. The images are quantized so coarsely in order to make the quantization artifacts as visible as possible for demonstration purposes. If we were to quantize an original source to 8-bits, as would be more normal in digital imaging, we would only need 2 brightness levels of headroom on either side of the image (i.e. the brightness levels of the original image could be anywhere between 2 and 253). This is a much more acceptable limitation.

2.1 Quantization of Images

The quantization of a greyscale image is a scalar problem, just as it is in the case of audio. Thus the theory of quantization needs no adjustment when applied to monochrome images. Since this mathematical theory has already been developed, we will not repeat it here. Instead, this and subsequent subsections in this chapter will demonstrate the visual analogue of the phenomena we have already discussed in quantization and dither of audio signals [48].

As in the case of an audio signal, if the quantization step is not small enough, then there are going to be very visible errors in the quantized signal. Figure 2.5 demonstrates this effect. Notice the sharp contrasts in brightness, especially in the slow gradient in the background, as the brightness values make a transition to a new quantizer level. Figure 2.6 shows the error between this quantized signal and the original image, Figure 2.4, multiplied by a gain factor of

1.4 in order to make the artifacts slightly more visible. This gain factor has been applied to all error signals presented in this chapter. The error signal in this case is visibly dependent on the input signal and is clearly correlated with it.

The first examples of the quantization of images that we have found in the literature comes from Schroeder [68], who studied the problem for use with microfilm plotters. This paper is also the first example we have found of dithered quantization being used on images.

2.2 Subtractive Dithering of Images

Following the development of our audio counterpart, the first step to removing the unwanted artifacts in the quantized image is to try subtractive dither. The theory has not changed at all, so we will be adding an RPDF dither to brightness values of the image, quantizing to 3-bits, and then subtracting the same dither signal.

The result of this procedure can be seen in Figure 2.7, and the total error of the output can be seen in Figure 2.8. As in the audio case, subtractive dither represents the best-case scenario; the error of the signal is an additive benign white noise that is statistically independent of the input, and this is accomplished with the least amount of residual noise power possible.

2.3 Nonsubtractive Dithering of Images

Just as in audio applications, subtractive dither is not practical since it requires synchronization of a dither signal on opposite ends of a media channel. So we turn to nonsubtractive dither.

First, Figure 2.9 shows the result of a nonsubtractive dither scheme when using RPDF dither. Of specific interest in this figure is the brightness gradient in the background. Notice that as the brightness slowly darkens, the amount of noise present seems to vary periodically. This is the same effect we observed in the audio case. The error signal, Figure 2.10, confirms that the error has an average value of zero, but that the error power varies with the input signal. This is noise modulation. Notice, however, that the blatant distortions seen in Figure 2.5 are eliminated. The error becomes a white noise with zero-mean, but its variance depends on the local brightness amplitude of the sample points. In fact, we can still make out some features of the original image from this error signal. Therefore, we can conclude that, like the human ear, the human eye is sensitive to the second moment of the error signal.

Next, we will apply TPDF dither to the brightness level of the original signal. The result can be found in Figure 2.11. The error signal can be found in Figure 2.12. The variance of the error is now constant across the entire image. Note that the noise is much more visible in Figure 2.11 than it is in the subtractive dither case, Figure 2.7. This is because the variance of the error is three times as much as it is in the subtractive dither case; $\frac{\Delta^2}{4}$ compared to $\frac{\Delta^2}{12}$.

It is worth mentioning that we used TPDF dither in the case of digital audio because tests show that the human ear is insensitive to statistical moments higher than the second. However, it has been suggested by Brinton [11] that the human eye does have some sensitivity to the statistical third moments. At this point we will take a moment to investigate this possibility.

For a nonsubtractively dithered system using a midriser-type quantizer and TPDF dither, the third moment of the error is shown in Figure 2.1 [73]. The third moment of the error has a peak positive value at about 0.786Δ, and a peak negative value at about 0.214Δ.

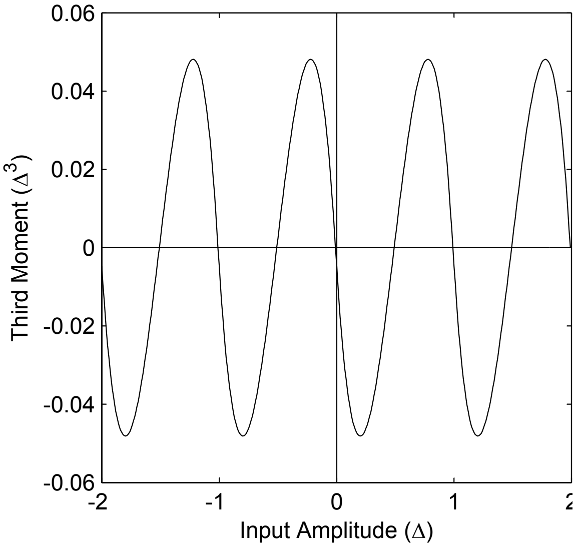


Figure 2.1: Third moment of the error of a TPDF nonsubtractively dithered system.

Figure 2.13 shows a monochrome image that was specifically designed to try to display the third moment of the error signal. The upper left corner of the outer square, and the lower right of the inner square portions have a brightness value at a quantum level. The third moment of the error is zero for these portions of the image. The lower right of the outer square, and the upper left of the inner square portions have a brightness value that is halfway between quantum steps, which also has a zero third error moment. The outer upper right and inner bottom left portions have a brightness value that gives the third moment a peak negative value. The remaining two portions give the third moment of the error a peak positive value. The resulting quantized image, shown in Figure 2.14, shows the result of quantizing the image when using TPDF dither. The use of TPDF dither means that the first two statistical moments of the error signal are controlled. Therefore, any remaining visible artifacts in the error signal should be due to the third (or higher) statistical moments. The error signal is shown in Figure 2.15. We have found, when viewed from a suitable distance so that one is unable to distinguish between the individual sample points (in order to satisfy the reconstruction operation of the Sampling Theorem), that the error image looks like a uniformly white noise. Thus, we conclude that the human eye is insensitive to statistical moments higher than the second, and will continue our theory with this assumption.

Brinton's [11] statement that the third moment of the error might be visible can be attributed to a couple of different factors. Since we were unable to procure a quality facsimile of the images used in Brinton's thesis, we can only conjecture at what the causes might have been for his claim. The most probable causes might have been clipping of brightness levels in the error images, or Brinton may have been studying the error images too closely, thus violating the reconstruction aspect of the Sampling Theorem. If the brightness levels of the images have been clipped, or any other significant non-linear distortion occurs, then the reproduction of the image can convert higher statistical moments into first and second moments, thus giving the false appearance of sensitivity to more than the first two statistical moments of the error.

As far as the author is aware, Brinton [11] is the first, and possibly only, author to study the effects of TPDF dither on images, and also the only other author to contemplate the error moments caused by quantization of an image. It is important to note that the addition of dither to an image signal is also referred to as "threshold modulation" in the literature. Billotet-Hoffmann and Bryngdahl [9] are sometimes attributed [42] with the discovery of dithering images, but, as mentioned earlier, Schroeder [68] is the first source that we have been able to find on the matter.

2.4 Noise Shaping of Images

The process of noise shaping in digital image processing is often referred to as "Error Diffusion". The technique was first published by Floyd and Steinberg [24], but the study of 2-D recursive filtering goes further back to the work done primarily by Shanks [36, 77, 70]. Shanks was interested in the properties of 2-D recursive filters for applications in geophysics. Most of the early theory on noise-shaping of discrete two-dimensional signals was published by Shanks, and we will investigate this theory in the following chapter.

Just as in the audio case, the process of noise shaping of images involves distributing the quantization error at a sample point to neighbouring sample points that have not yet been processed. If we are quantizing an image in a left-to-right and then top-to-bottom manner, the unprocessed sample points will be those to the right of the current sample, and anything in rows lower than the current sample. The following diagram illustrates this, where the 'o' represents a processed sample, an 'x' represents an unprocessed sample, and a '$\cdot$' represents the sample point being processed.

$$\begin{array}{ccccccc}
o & o & o & o & o & o & o \\
o & o & o & o & o & o & o \\
o & o & o & \cdot & x & x & x \\
x & x & x & x & x & x & x \\
x & x & x & x & x & x & x
\end{array}$$

The error may be distributed from the current sample point to any of the x's shown. In this fashion, we can shape the spectrum of the total error of the final signal using the noise transfer function $1 - H(z_1, z_2)$.

To contrast the audibility curve, Figure 1.19, the sensitivity of the human eye to brightness changes as a function of spatial frequency is shown in Figure 2.2. This particular visibility curve comes from the work of Mannos and Sakrison [49]. Just as with the human ear, the human eye is most sensitive to low frequency stimuli, with a peak in sensitivity at approximately 8 cycles per degree. Of course, this curve is an average, and individual experiences may differ. It is important to note that spatial frequency varies with viewing distance, becoming higher frequency as we move farther away from the image.

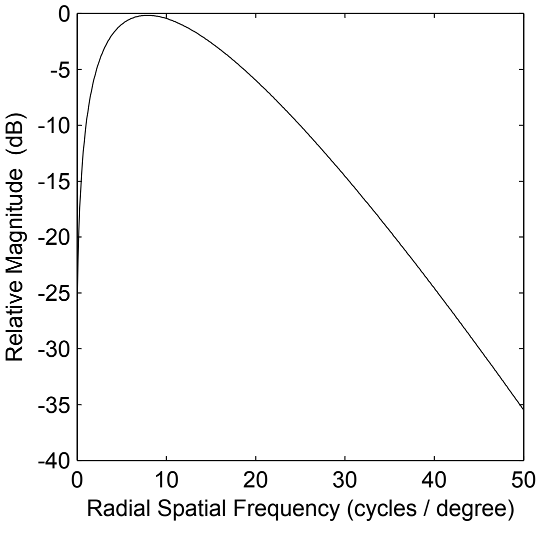


Figure 2.2: Sensitivity of the human eye to low-contrast as a function of frequency.

A search through the literature on this topic has shown that most applications of noise shaping, or error diffusion, on images has been regarding the halftoning problem. The halftoning problem is the question of how to give the illusion of a continuous brightness scale given only two colours. The two colours are usually black and white. This is equivalent to quantizing an original image to only 1-bit. In these instances, there is not enough headroom in the quantizer to allow us to apply the full theory of dithered noise shaping. As a consequence, we have found no mention of TPDF dither applied to images in the literature. In fact, unless otherwise stated, it is generally assumed that no dither is used at all in the process of error diffusion. However, multi-bit rendering devices are now common (computer monitors, television sets, and even newer inkjet printers), and so properly dithered noise shaping does have applications.

The filter given by Floyd and Steinberg [24] in their original paper was

$$1 - H(z_1, z_2) = 1 - \frac{7}{16} z_1^{-1} - \frac{3}{16} z_1 z_2^{-1} - \frac{5}{16} z_2^{-1} - \frac{1}{16} z_1^{-1} z_2^{-1}$$

This means that the quantization error at a given pixel was distributed to the neighboring sample points in the following amounts

$$\begin{array}{ccc} & \times & \frac{7}{16} \\ \frac{3}{16} & \frac{5}{16} & \frac{1}{16} \end{array}$$

The × denotes the current sample point. This design has become a standard benchmark in the literature. The log-magnitude of this noise shaper design is shown in Figure 2.3. If we think of a radial branch of this surface, it is a poor fit to our inverted visibility curve. Later, we will try to improve on this design. Notice that the noise transfer function is zero when $z_1 = z_2 = 1$. This is responsible for the null at $(0, 0)$, or "dc", in Figure 2.3.

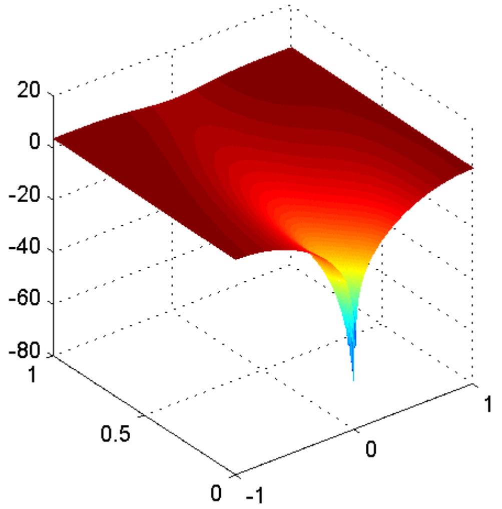


Figure 2.3: Floyd and Steinberg log-magnitude (dB) plotted against spatial frequency. The surface is even, and so only half of the surface is shown here.

As in the case of audio signals, we need to properly dither the image before quantizing in order to ensure that we are controlling the total error properly. Figure 2.16 shows the result of using the Floyd and Steinberg noise shaper on an image without using any dither. Notice that there are undesirable artifacts left over, such as noise modulation and patterns in the error signal due to limit cycle oscillations in the error feedback loop. Also, it was found that undithered noise shaping produces an artificial edge enhancement, making the image seem sharper than it actually is [40]. The error signal can be found in Figure 2.17.

If using proper TPDF dither with the Floyd and Steinberg noise shaper, one produces images such as Figure 2.18. Notice that the undesirable artifacts from undithered noise shaping have disappeared. The error of this image manifests itself as a benign background noise. The error can be seen in Figure 2.19. Figure 2.18 is actually much noisier than Figure 2.7, the subtractively dithered image, although it appears less noisy when viewed from an appropriate distance. Since the lower frequencies of the error signal are suppressed in the noise shaped image, Figure 2.18 actually looks preferable to Figure 2.7 when viewed from a reasonable distance (i.e. far enough from the image so that one cannot distinguish individual sample points).

It is important to note that for the statistical properties of noise shaping to take effect, we require that the size of the noise shaper be small in comparison to the input. Otherwise, the output will contain artifacts related to the fact that all of the buffers of the system have not had a chance to affect the input. This is important when working with very small images. Also, to avoid artifacts around the perimeter of an image, it is advised to create a set of buffer sample points by extending the image across each of the boundary edges of the image.

As mentioned before, when error diffusion is mentioned in the literature, undithered noise shaping is meant. In fact, very few authors have attempted to dither the noise shaping system as we have, and, to our knowledge, nobody has tried using TPDF dither. Rectangular pdf dithers of less than 1 LSB in width have been tried, and have offered some improvements [80], but, given the 1-D theory, we know that these dither signals are insufficient in order to get rid of all the undesirable artifacts in the final image. To add to the confusion, the term "dither" has been given a different, non-equivalent definition in image processing. In image processing, the term dither usually refers to a mask pattern. These are two things to be aware of when reading through image processing literature.

2.5 Dithered Quantization of Colour Images

For colour images, the quantization procedure becomes a vector process. In this section, we will illustrate the properties of applying dithered quantization and noise shaping on colour images in a direct way. We perform the dithered quantization and noise shaping procedure on each of the three colour components, red, green, and blue, individually, and then recombine these components to form the final quantized image. This is not an optimal method of quantizing colour images.

Figure 2.20 shows an original 24-bit image, 8-bits per red, green, and blue component of the image. Figure 2.21 shows subtractive dithering of the original when we use a different dither signal for each of the red, green, and blue components and quantize each of these components individually to 3-bits. The result shown is the reconstituted image out of these three components. Figure 2.22 shows the results of nonsubtractive dithering using TPDF dither. Again, a different dither signal is used for the red, green, and blue components, and each component is quantized to 3-bits. Figure 2.23 shows the original when quantized to 3-bits per colour component using TPDF dither and the Floyd and Steinberg noise shaper design.

These last few images are purely illustrative. Dithered quantization and noise shaping of colour images is an area that could be researched further. The procedure outlined in this section is not the optimal approach to the quantization of colour images, merely a suggestion of how one might start to look at this theory.

2.6 Images

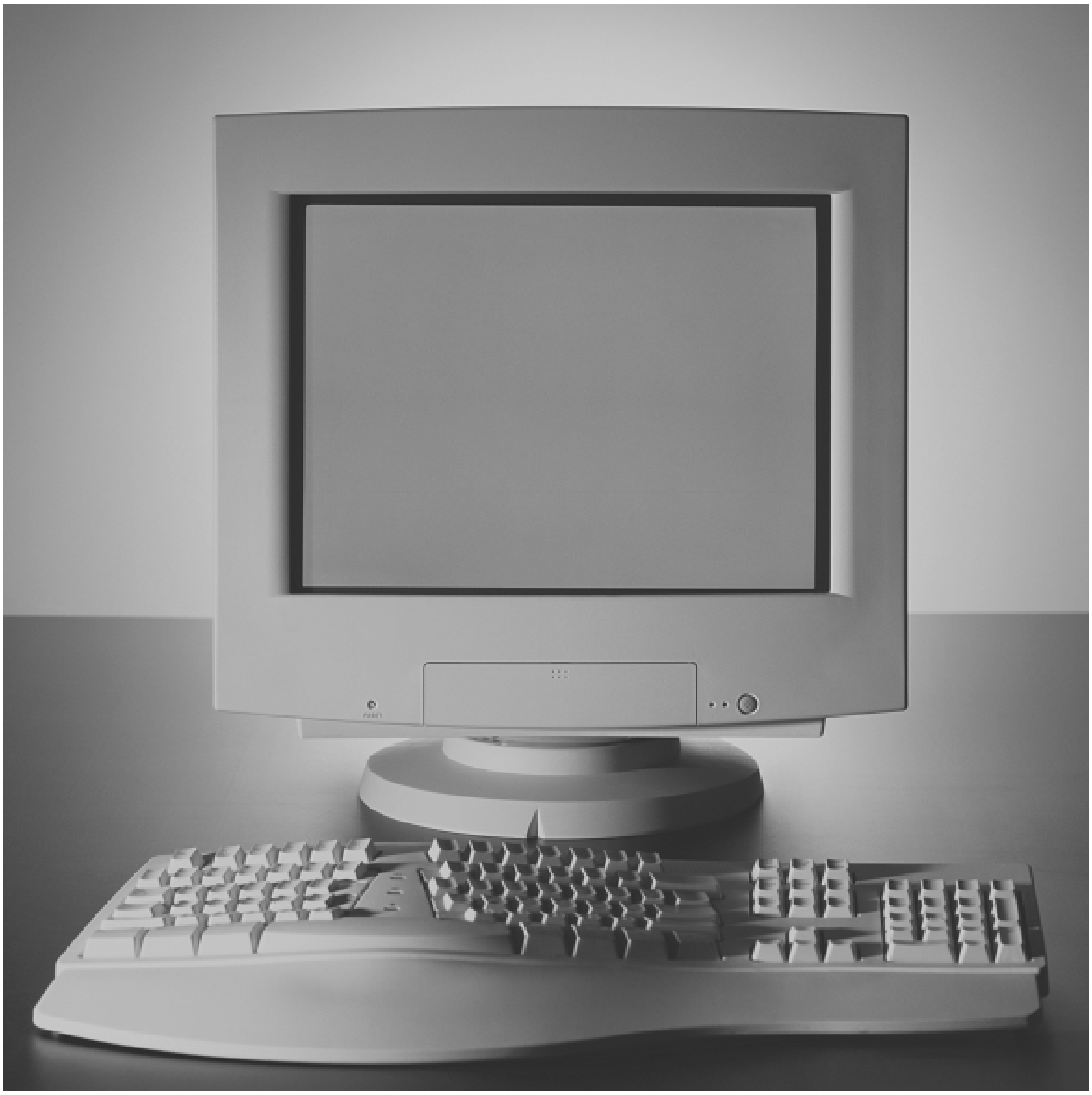


Figure 2.4: Original 585 × 585 image, 8-bit greyscale.

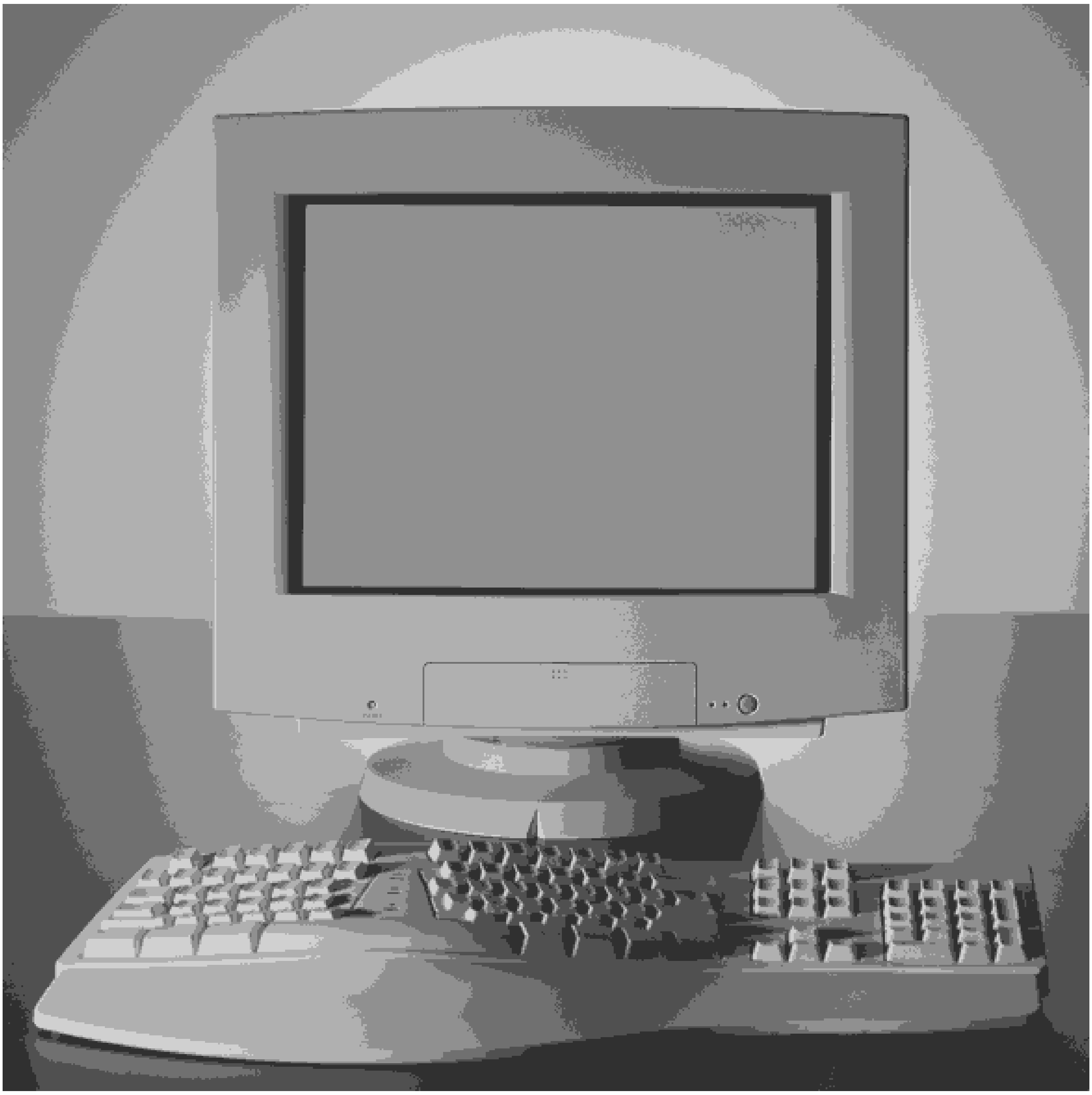


Figure 2.5: Original image quantized to 3-bits.

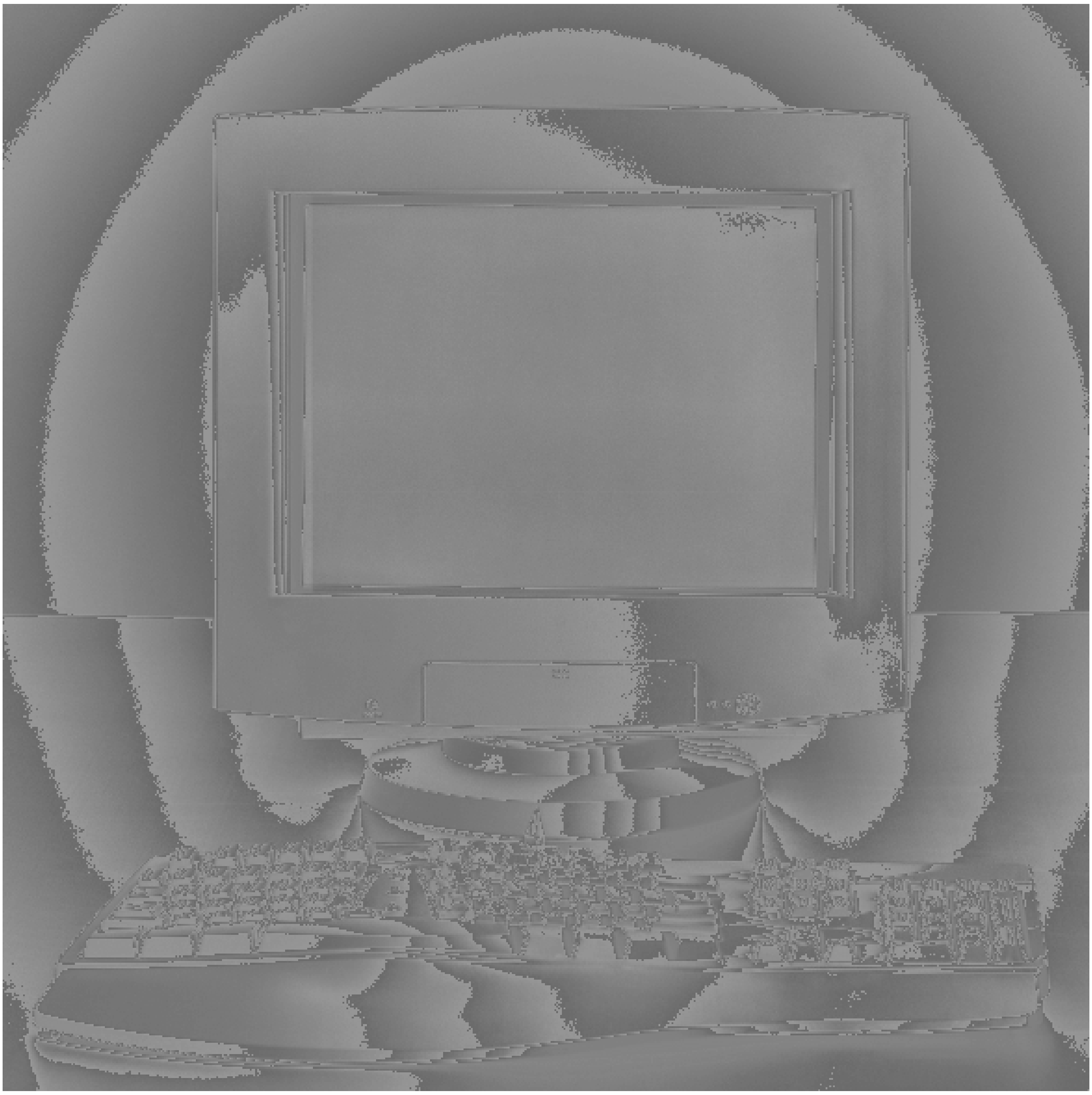


Figure 2.6: Error between the image quantized to 3-bits and the original image.

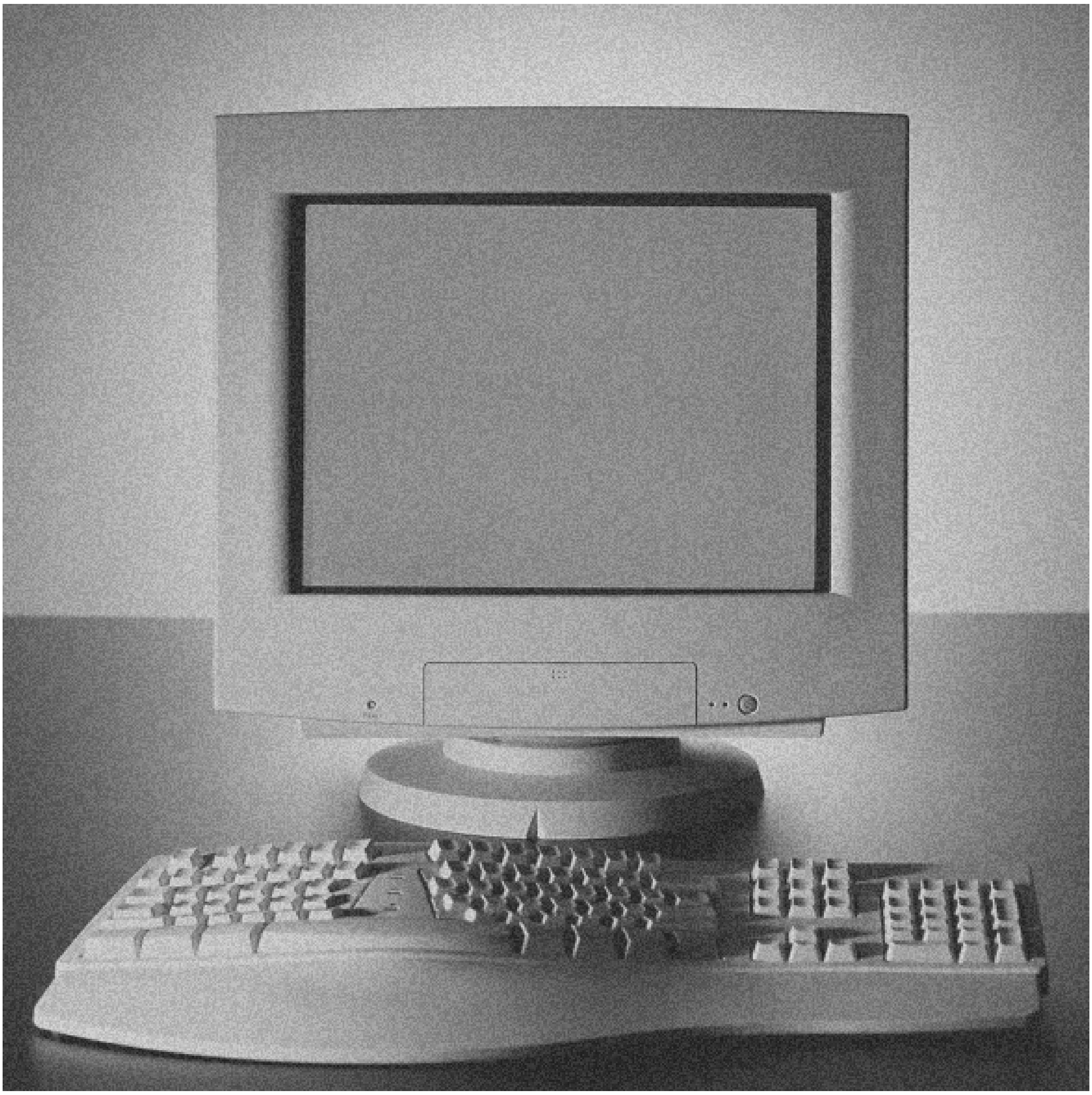


Figure 2.7: Original image quantized to 3-bits using a subtractive dither scheme and RPDF dither.

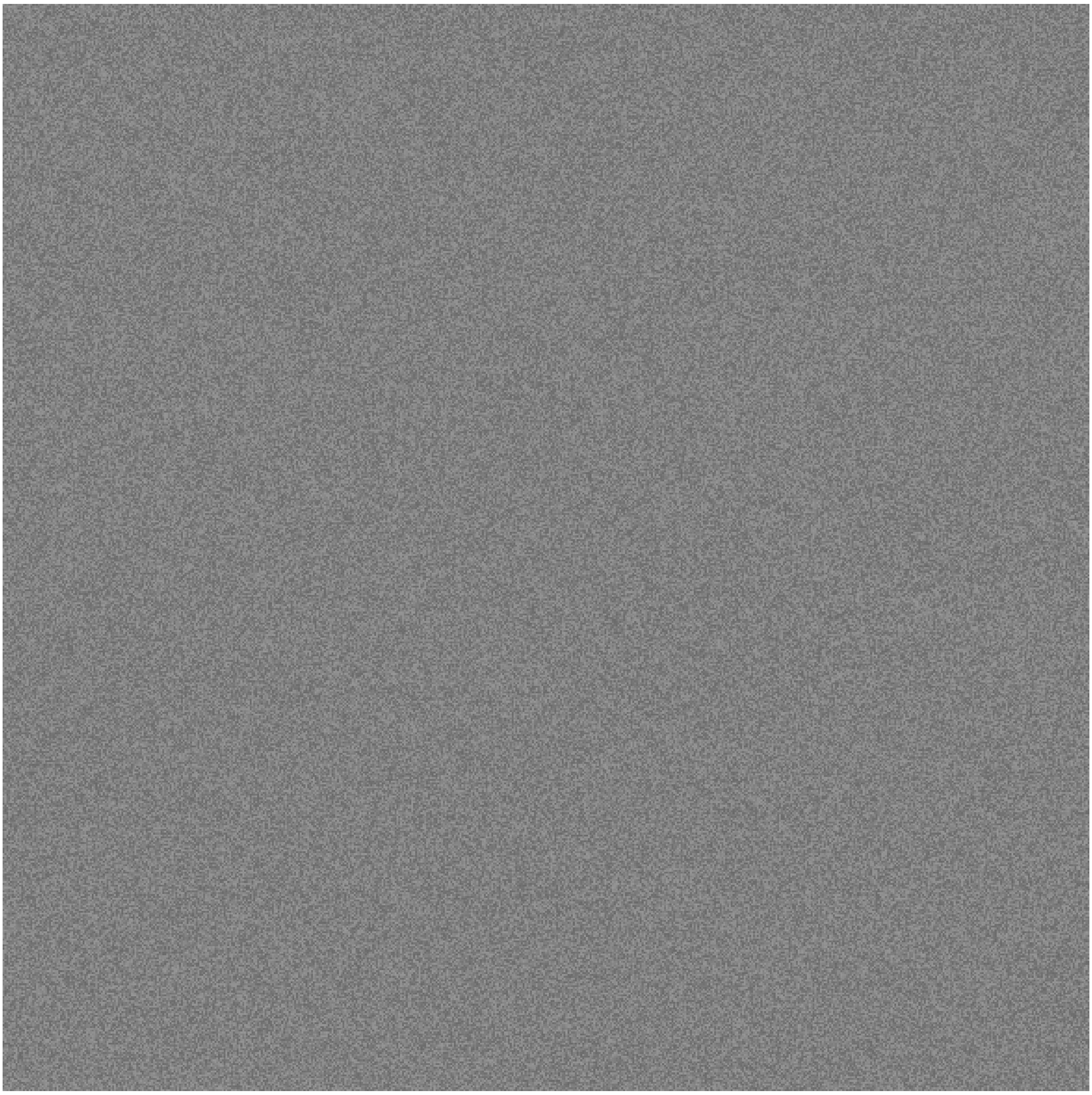


Figure 2.8: Error between the subtractively dithered image and the original image

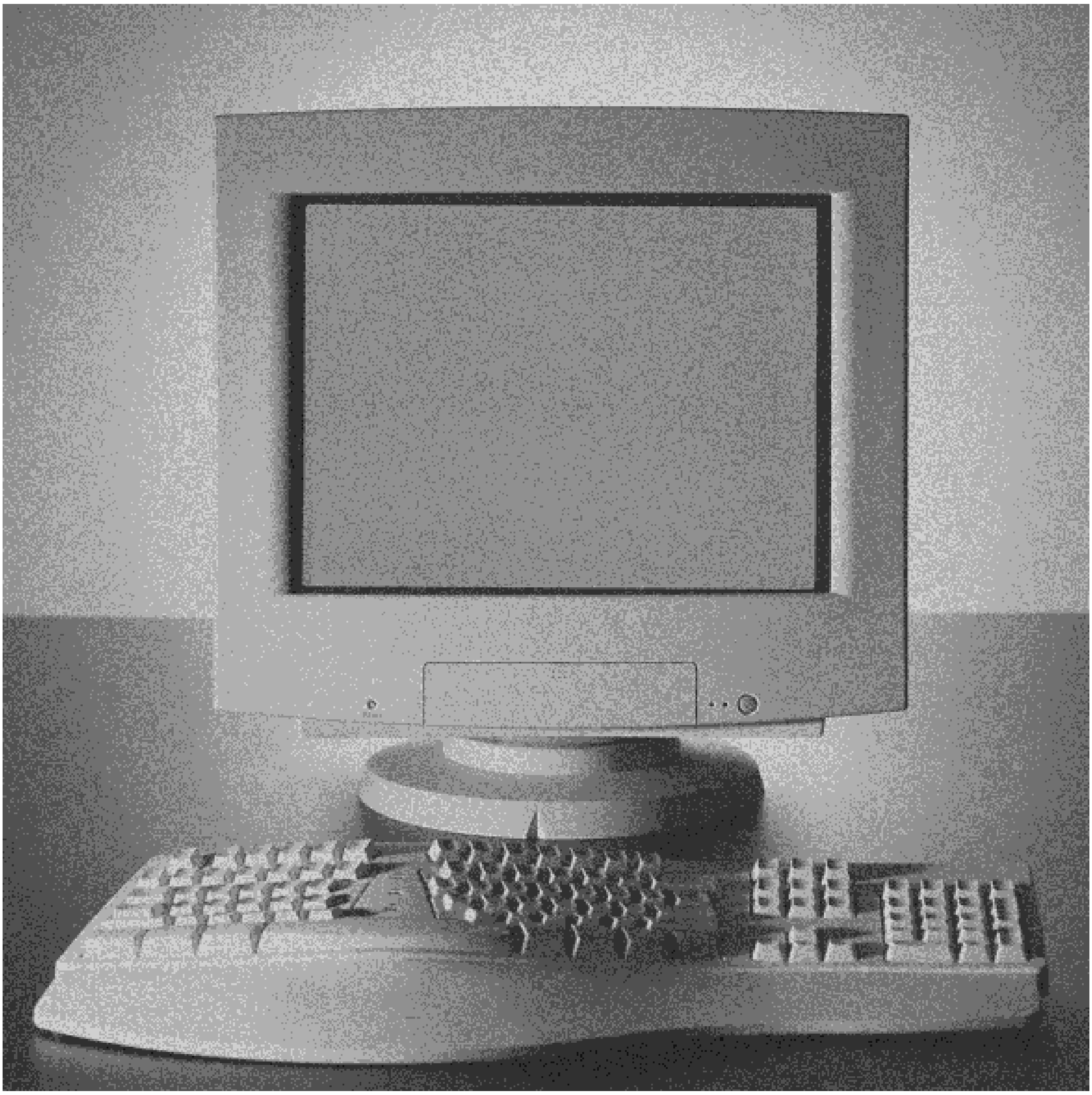


Figure 2.9: Original image quantized to 3-bits using a nonsubtractive dither scheme and RPDF dither.

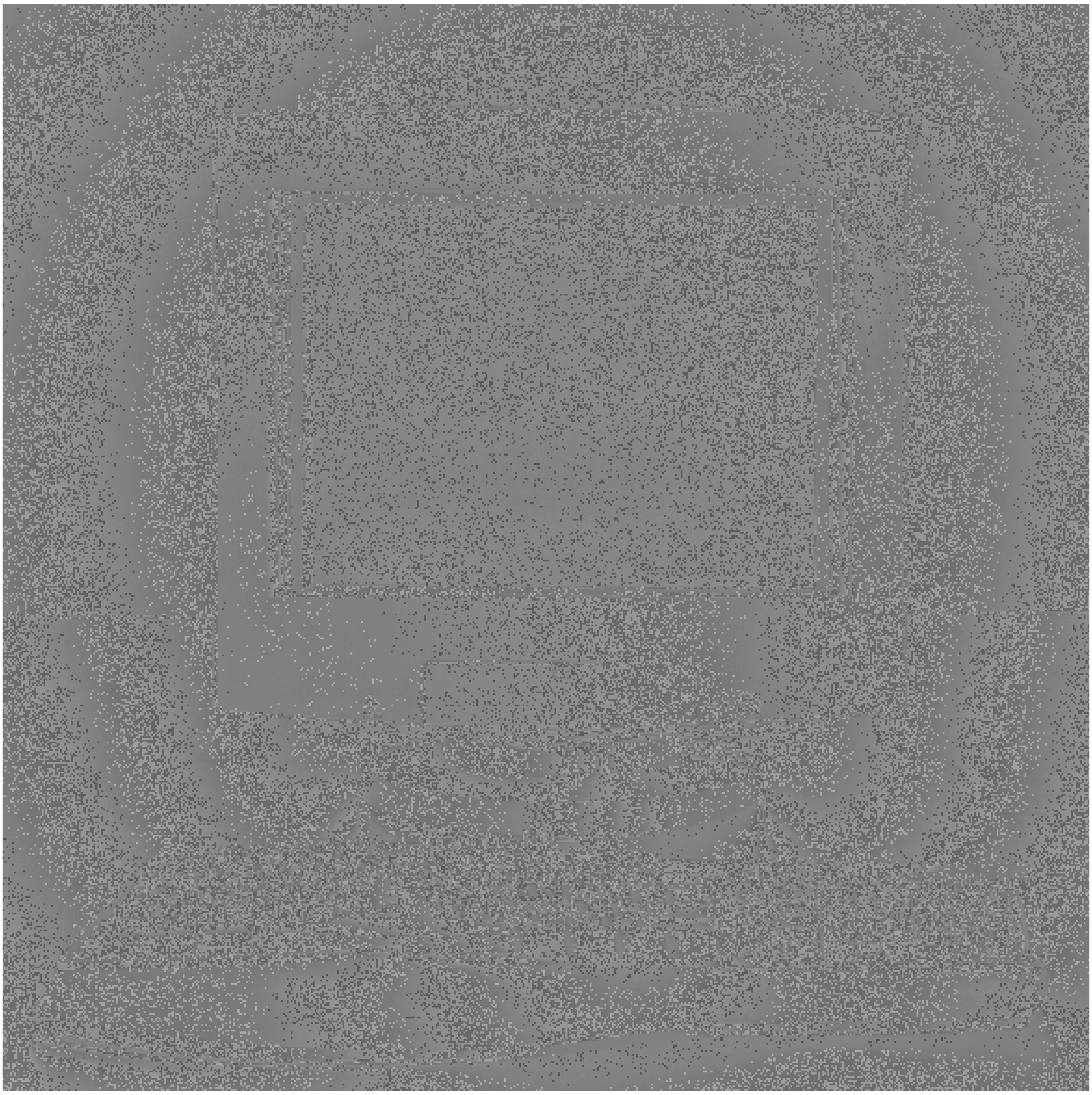


Figure 2.10: Error between the RPDF nonsubtractively dithered image and the original image.

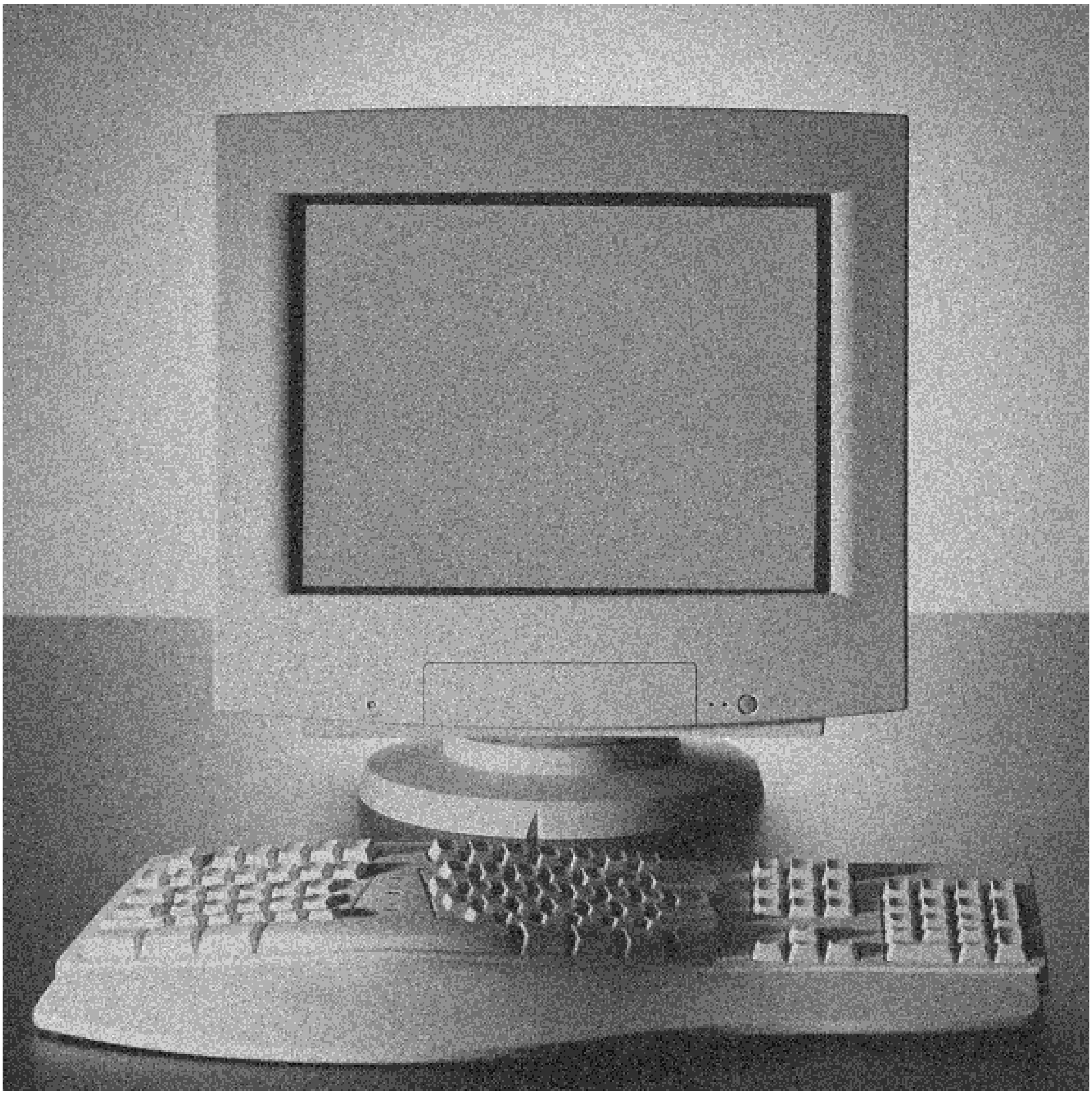


Figure 2.11: Original image quantized to 3-bits using a nonsubtractive dither scheme and TPDF dither.

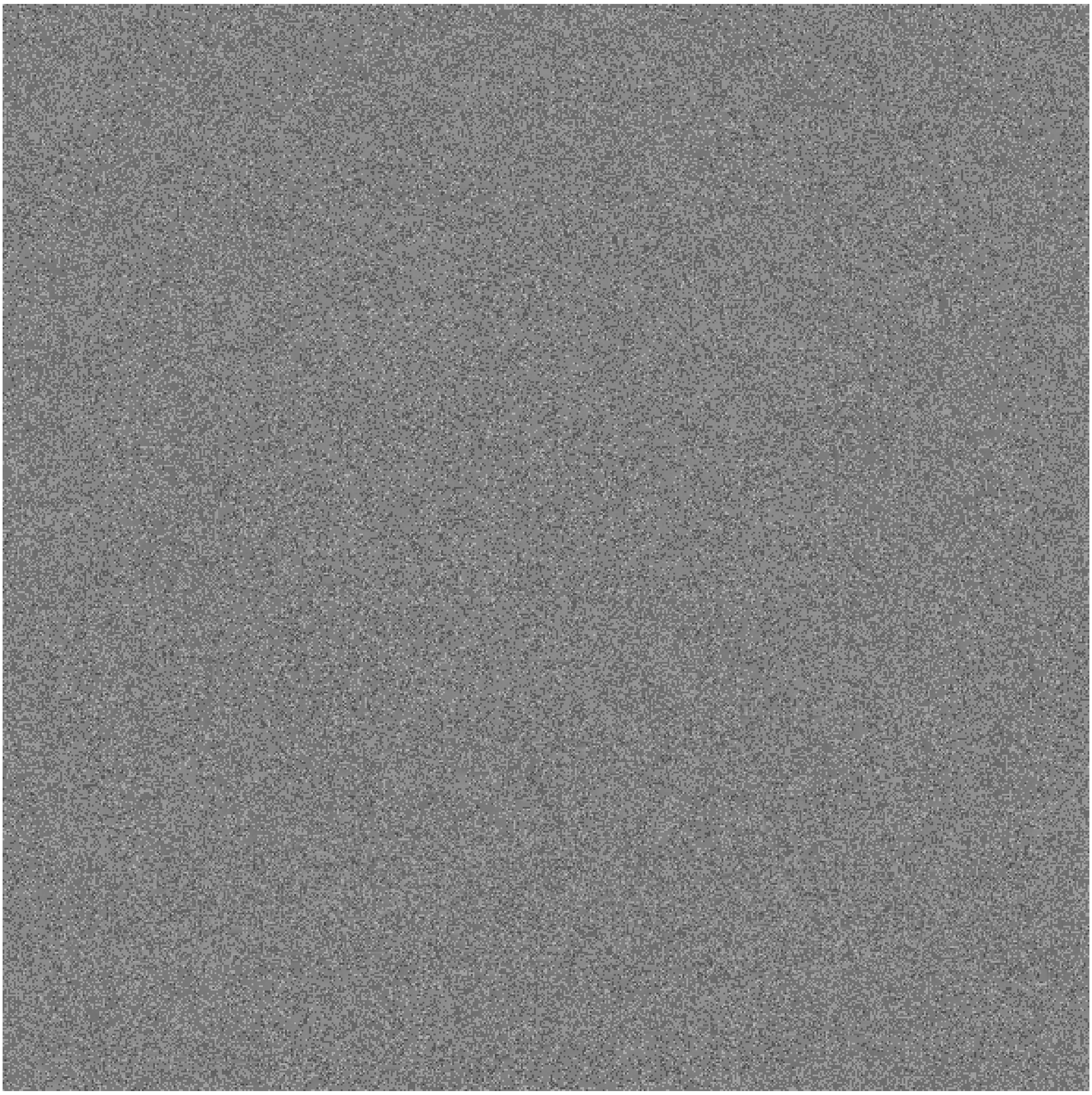


Figure 2.12: Error between the TPDF nonsubtractively dithered image and the original image.

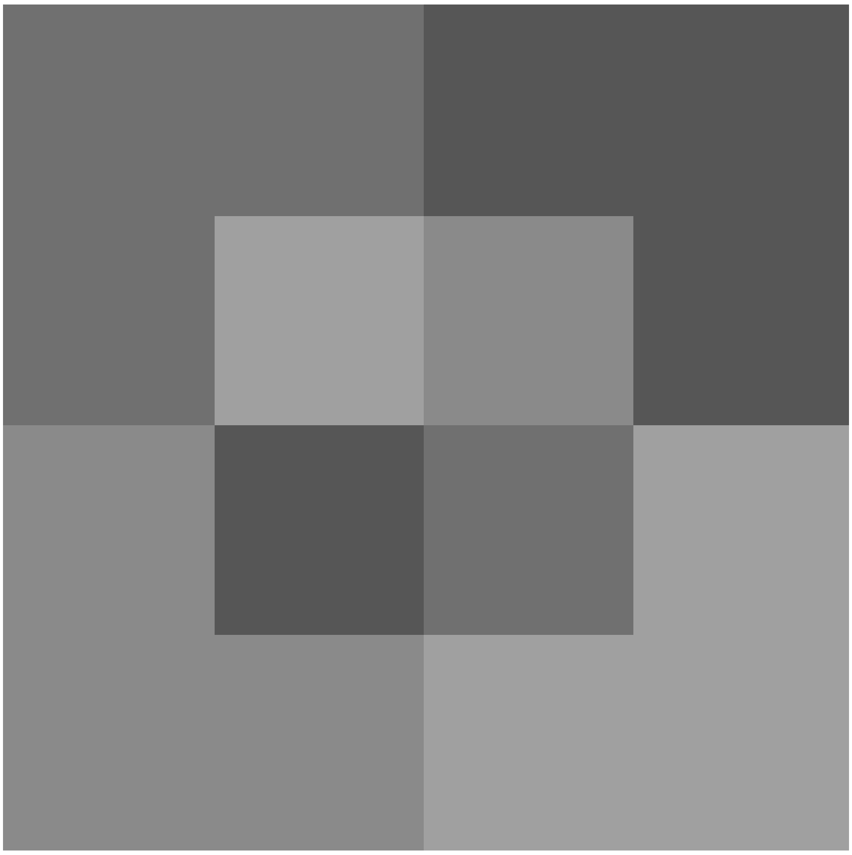


Figure 2.13: Original image used to try to observe the third moment of the error.

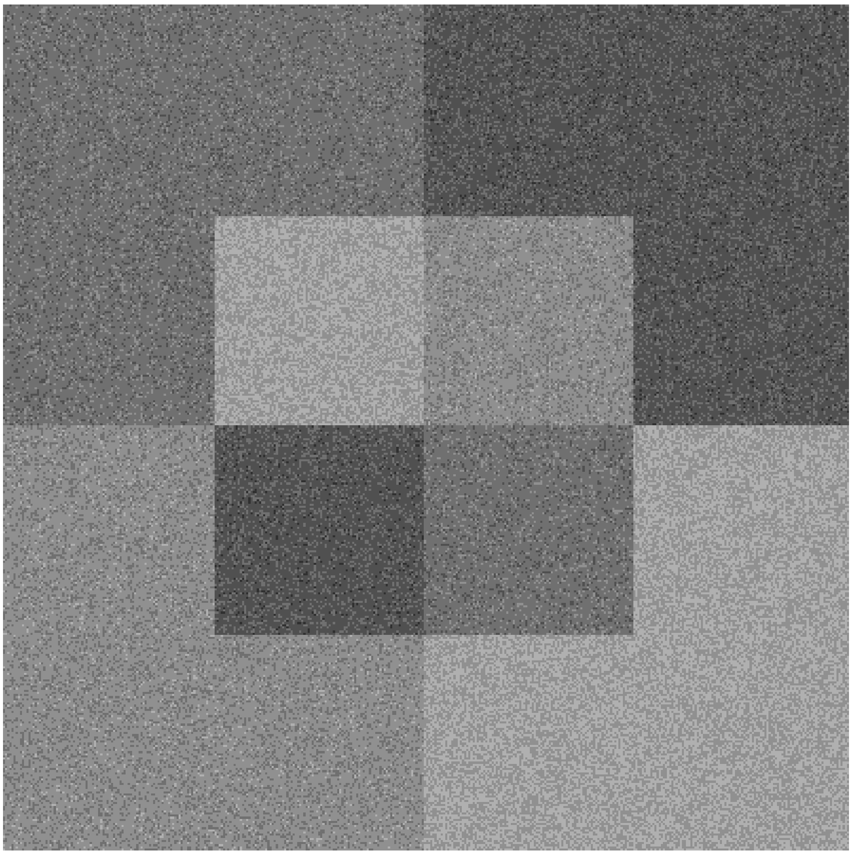


Figure 2.14: Quantized image using a nonsubtractive dither scheme and TPDF dither.

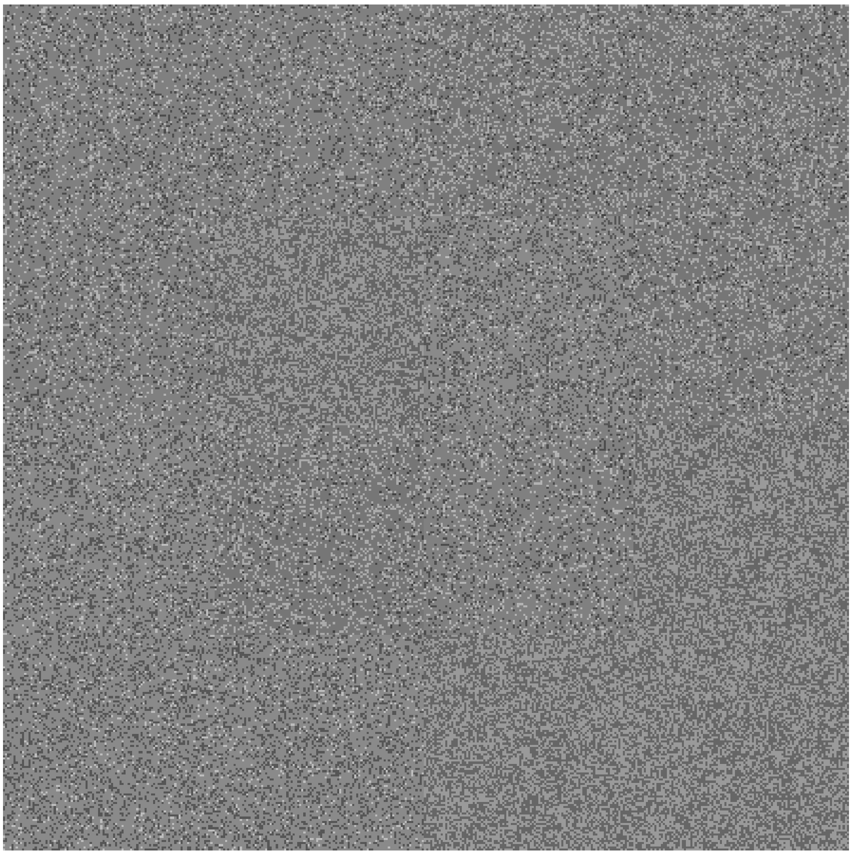


Figure 2.15: Error of the image attempting to display third error moments.

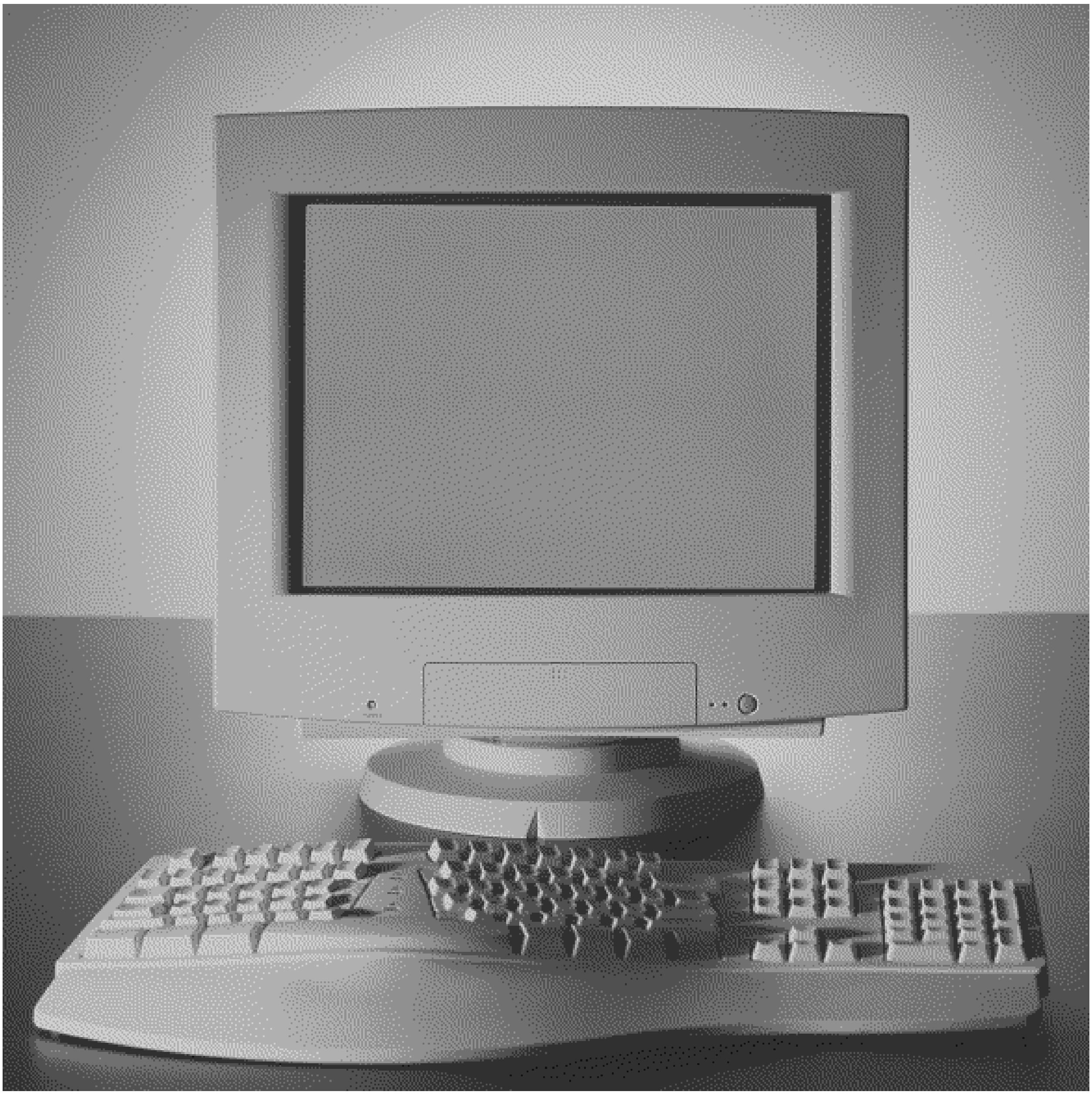


Figure 2.16: Original image quantized to 3-bits using a noise shaping scheme of Floyd and Steinberg's design with no dither.

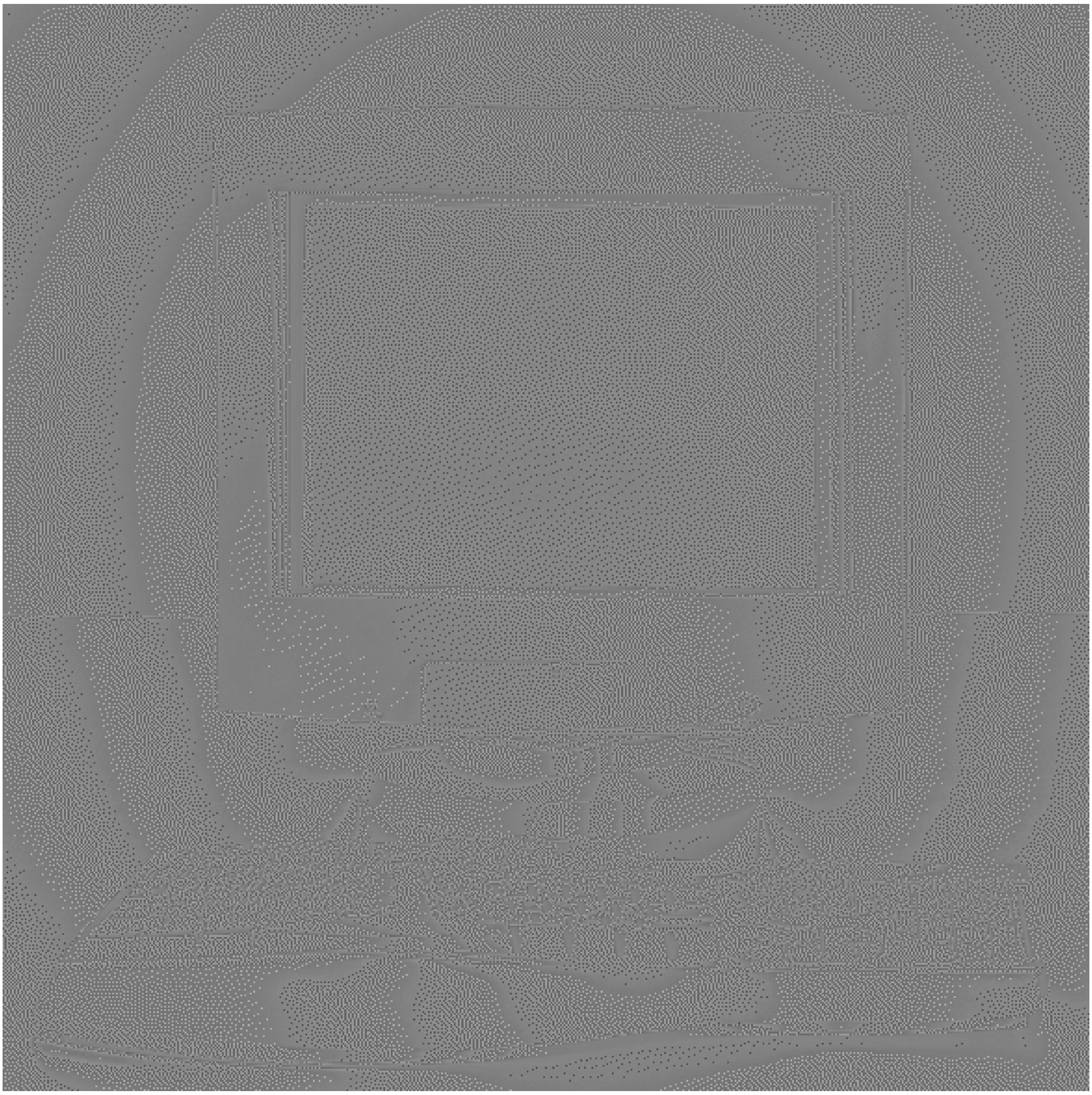


Figure 2.17: Error between noise shaped imaged and the original image.

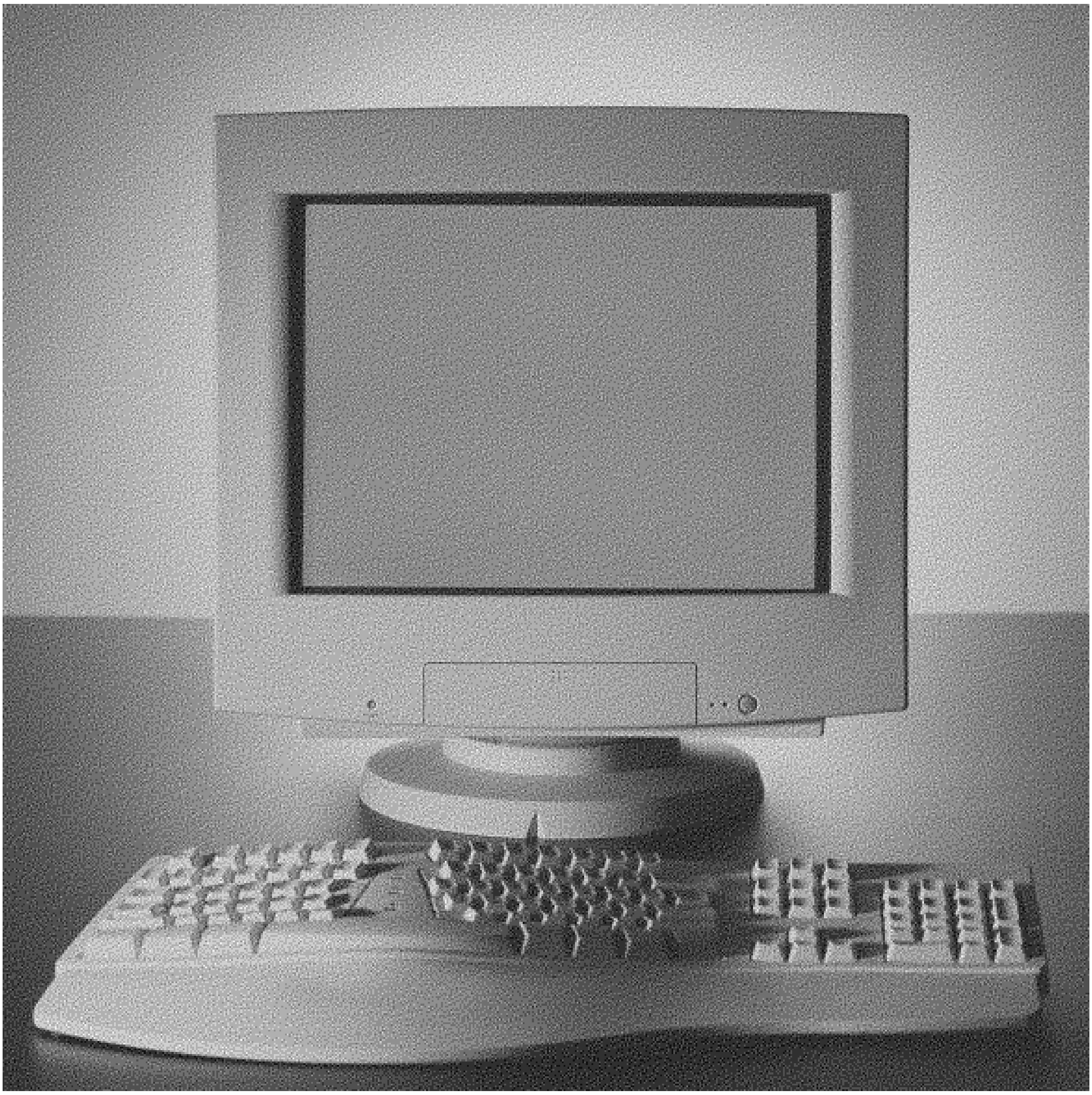


Figure 2.18: Original image quantized to 3-bits using a noise shaping scheme of Floyd and Steinberg's design with TPDF dither.

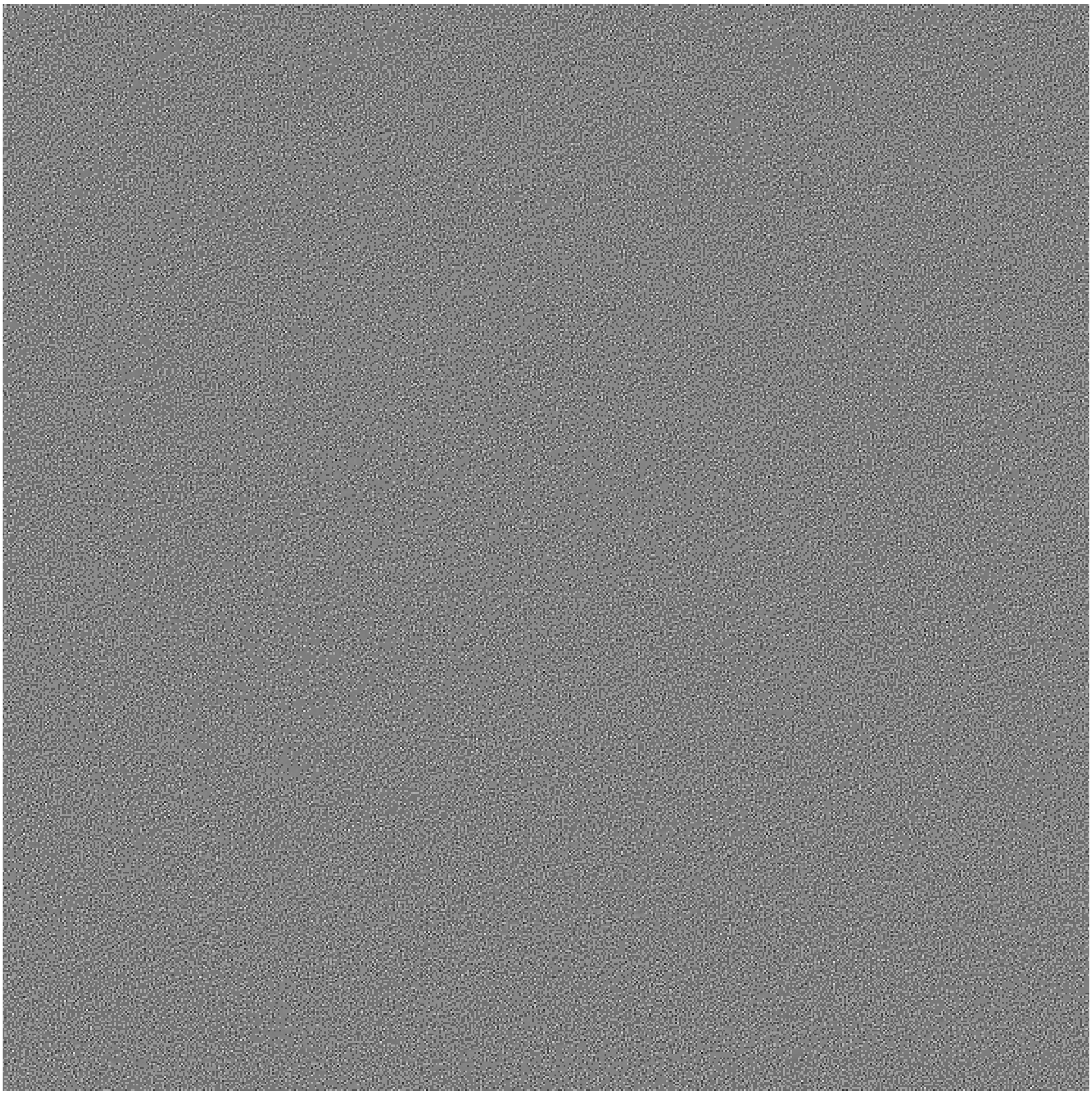


Figure 2.19: Error between noise shaped image and the original image.

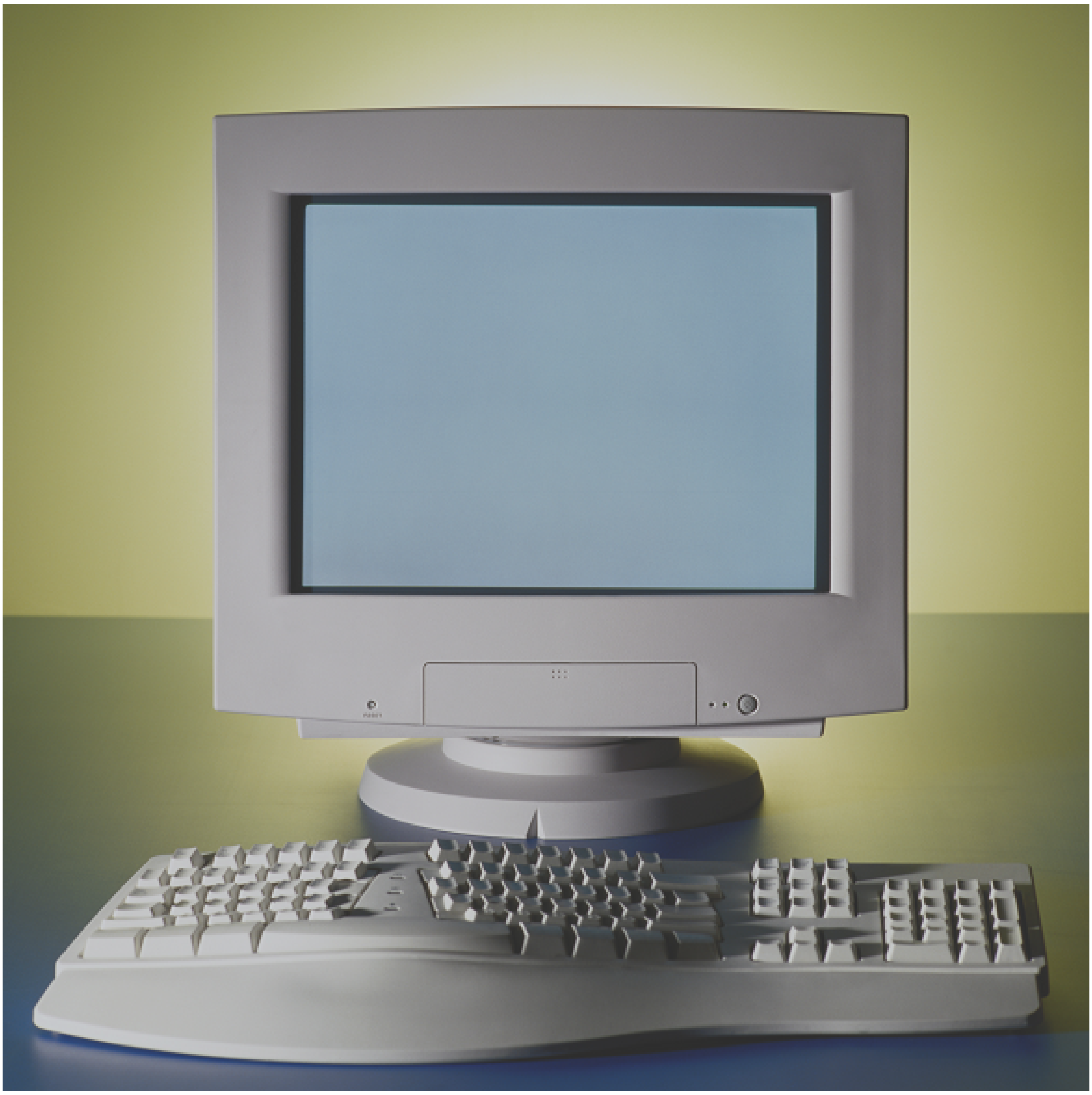


Figure 2.20: Original 24-bit colour image.

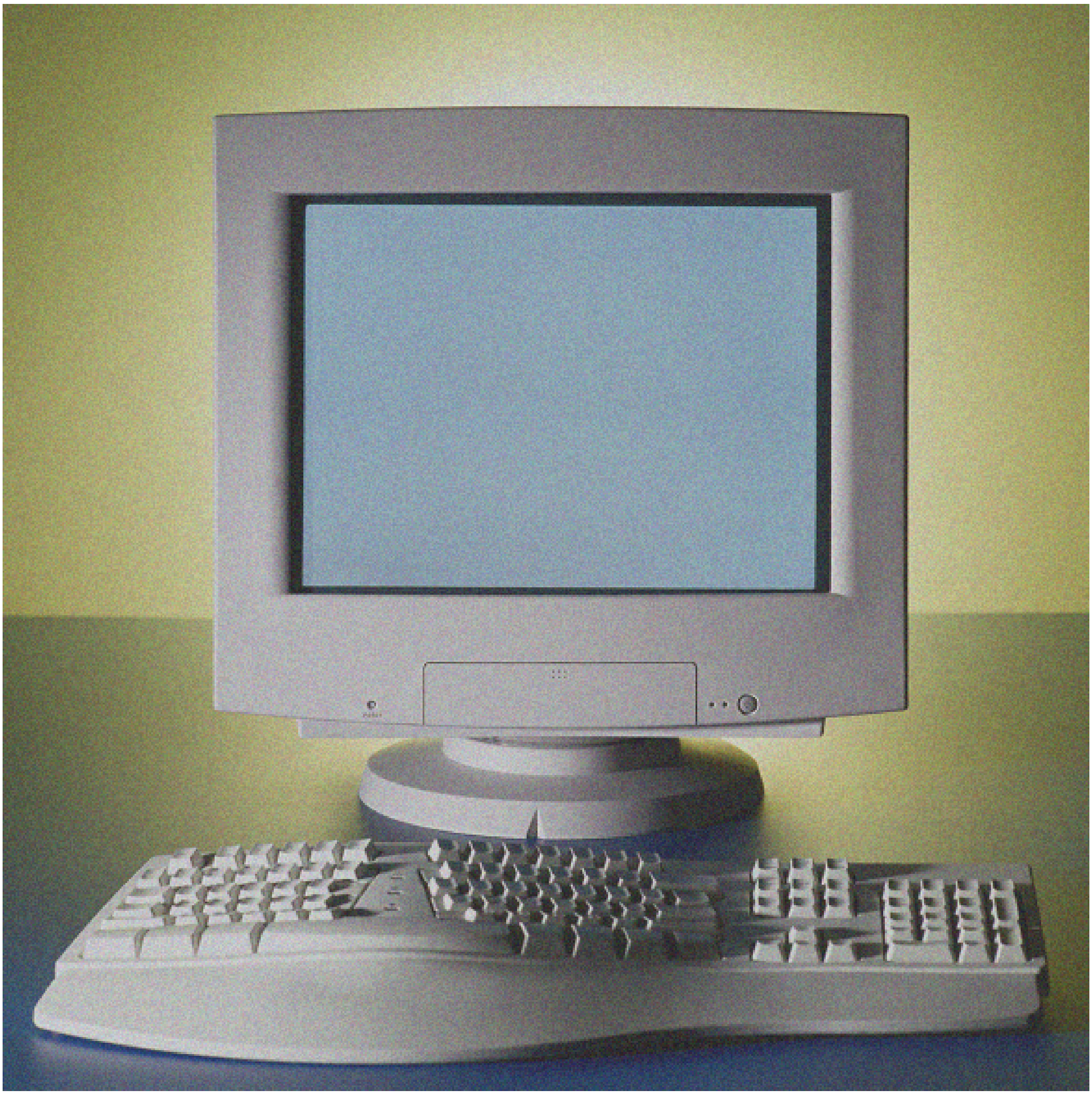


Figure 2.21: Colour image quantized to 3-bits per colour using the subtractive dithering scheme with RPDF dither.

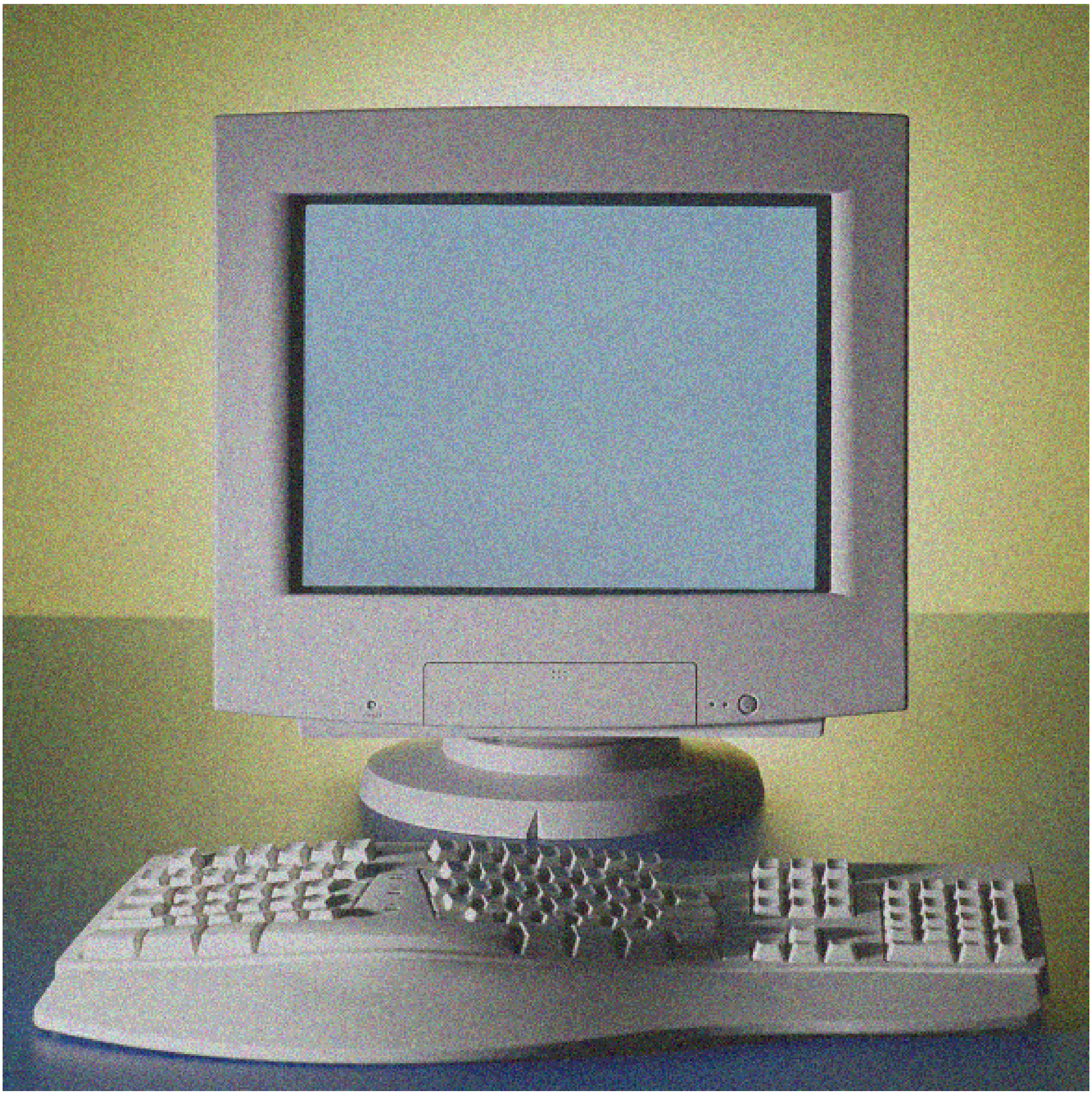


Figure 2.22: Colour image quantized to 3-bits per colour using the nonsubtractive dithering scheme with TPDF dither.

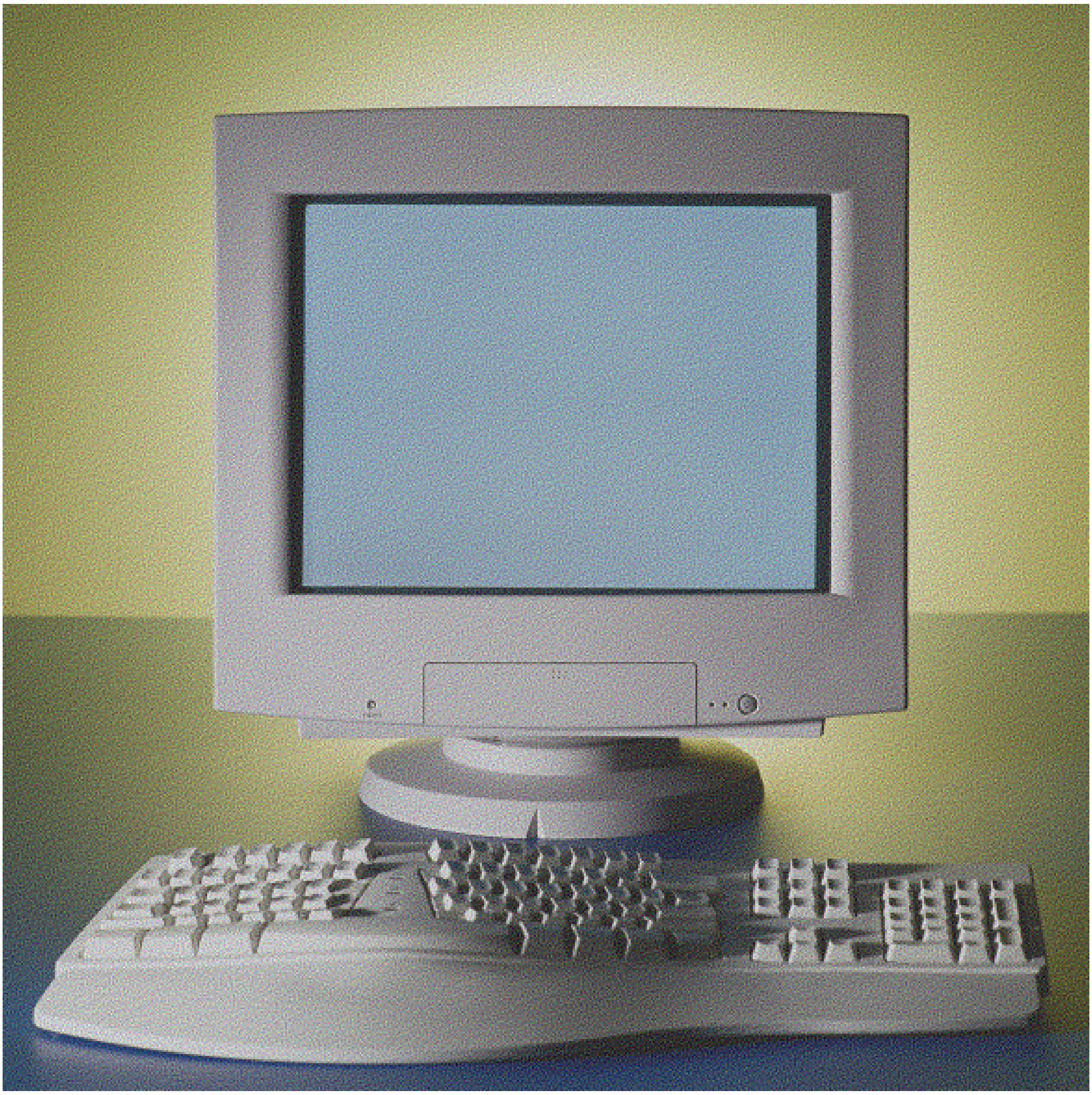


Figure 2.23: Colour image quantized to 3-bits per colour using TPDF dither and the Floyd and Steinberg noise shaper.

Chapter 3

Extending Jensen's Inequality

Jensen's inequality is crucial to the 1-D theory of noise shaping. It tells us that a minimum-phase system will give us the least amount of visible noise in the resultant signal, given that the frequency response of the noise must have a certain log-magnitude shape. In this chapter, we will work towards extending this concept to the case of 2-D signals.

To this end, we will need to start off with some preliminary material about 2-D sequences and systems. We will see that 2-D systems can be grouped into equivalence classes based on certain characteristics of the system, such as stability and volume under the log-magnitude surface. These equivalences will be exploited in our proof of an Extended Jensen's Inequality.

All sums in this chapter are assumed to be for all integer values of the summation index unless otherwise explicitly stated.

3.1 One-Sided Sequences

A 2-D sequence, $a(n_1, n_2)$, can be written out in long form as follows

$$\begin{array}{ccccc} & \vdots & & \vdots & \\ \cdots & a_{-1,-1} & a_{0,-1} & a_{1,-1} & \cdots \\ & a_{-1,0} & \times & a_{1,0} & \\ \cdots & a_{-1,1} & a_{0,1} & a_{1,1} & \cdots \\ & \vdots & & \vdots & \end{array} \tag{3.1}$$

The central × is the origin, and all numerical values of the sequence are listed outwards from here. To the right of the origin are the positive values of n_1. Below the origin are the positive values of n_2. If the sequence has a non-zero value at the origin, then that value will be subscripted with an $\times$. Often, we will explicitly mention which value is the origin so that there is no confusion whatsoever. Typically, only non-zero values are shown using this notation.

Example 3.1.1. Suppose that the only non-zero values of the sequence shown in Expression (3.1) are $a_{1,0}$ and $a_{-1,1}$. Then this sequence would be written as

$$\begin{array}{ccc} & \times & a_{1,0} \\ a_{-1,1} & & \end{array}$$

The × is still understood to be the origin. ♣

In 1-D, a one-sided sequence was defined as a sequence, $x(n)$, such that $x(n) = 0$ for all $n < 0$. So the sequence literally has values on only one side of the origin. In 2-D, the choice of which "side" of the origin to use becomes a little more arbitrary, but the principle will remain the same. Our definition of a one-sided 2-D sequence comes from Marzetta [52].

Definition 3.1.2. A 2-D one-sided sequence, $x(n_1, n_2)$, is any sequence that satisfies $x(n_1, n_2) = 0$ whenever $n_2 = 0$ and $n_1 < 0$, or $n_2 < 0$. ♦

For any linear relationship between our two indices, $n_1 = an_2$ (or $n_2 = bn_1$) for a non-zero integer a (or b), the sequence $x(n_1, n_2)$ reduces to a 1-D one-sided sequence.

Example 3.1.3. If we refer back to Equation (3.1), a one-sided sequence would look like

$$\begin{array}{ccccc} & & \times & a_{1,0} & \cdots \\ \cdots & a_{-1,1} & a_{0,1} & a_{1,1} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

♣

Example 3.1.4. Some examples of sequences that are not one-sided are the following:

$$\begin{array}{ccc} a_{-1,-1} & & a_{1,-1} \\ & \times & \\ a_{-1,1} & a_{0,1} & a_{1,1} \end{array}$$

This sequence is not one-sided because it has entries in the region $n_2 < 0$.

$$\begin{array}{ccc} a_{-1,0} & \times & a_{1,0} \\ & a_{0,1} & \end{array}$$

This sequence is not one-sided because it has an entry in the region $n_2 = 0, n_1 < 0$. Recall that the × denotes the origin. ♣

3.2 System Augmentations and Stability

A system is defined as a deterministic mapping defined on a suitable linear space of sequences . If T is a system and $x(n_1, n_2)$ is an input sequence, then $y(n_1, n_2) = T(x(n_1, n_2))$ is well-defined.

We are going to focus on a specific class of systems, referred to as linear shift-invariant systems. A linear system has the additional property that

$$T(a \cdot x_1(n_1,n_2) + x_2(n_1,n_2)) = a \cdot T(x_1(n_1,n_2)) + T(x_2(n_1,n_2)),$$

where a is a scalar. A system, T, is called shift-invariant if it has the property

$$T(x(n_1-m_1,n_2-m_2)) = y(n_1-m_1,n_2-m_2),$$

where $T(x(n_1,n_2)) = y(n_1,n_2)$ [46].

The class of linear shift-invariant systems has the property that its behaviour is completely determined by its response to an impulse signal. To see this, note that we can write

$$x(n_1,n_2) = \sum_{k_1}\sum_{k_2} x(k_1,k_2)\delta(n_1-k_1,n_2-k_2),$$

where

$$\delta(n_1,n_2) = \begin{cases} 1 \text{ if } n_1 = n_2 = 0 \\ 0 \text{ otherwise} \end{cases}$$

is called the Kronecker delta. If we use $x(n_1,n_2)$ as the input to the linear shift-invariant system T, where $h(n_1,n_2) = T(\delta(n_1,n_2))$ is summable, then

$$\begin{aligned} T(x(n_1,n_2)) &= T\left(\sum_{k_1}\sum_{k_2} x(k_1,k_2)\delta(n_1-k_1,n_2-k_2)\right) \\ \Rightarrow y(n_1,n_2) &= \sum_{k_1}\sum_{k_2} x(k_1,k_2)T(\delta(n_1-k_1,n_2-k_2)) \\ &= \sum_{k_1}\sum_{k_2} x(k_1,k_2)h(n_1-k_1,n_2-k_2). \end{aligned}$$

$h(n_1,n_2)$ is referred to as the impulse response of the linear shift-invariant system T, and entirely characterizes the system's behaviour to any input sequence $x(n_1,n_2)$. We can write this more compactly as a 2-D convolution:

$$y(n_1,n_2) = x(n_1,n_2) \overset{2}{*} h(n_1,n_2). \tag{3.2}$$

The inverse system cancels out the effect on the input. That is, $T^{-1}\left(T(x(n_1,n_2))\right) = T\left(T^{-1}(x(n_1,n_2))\right) = x(n_1,n_2)$. In particular, this is true when $x(n_1,n_2) = \delta(n_1,n_2)$. Under these conditions we have

$$\begin{aligned} T\left(T^{-1}(\delta(n_1,n_2))\right) &= T(h^{-1}(n_1,n_2)) \\ &= \sum_{k_1}\sum_{k_2} h^{-1}(k_1,k_2)h(n_1-k_1,n_2-k_2) \\ &= \delta(n_1,n_2). \end{aligned}$$

In particular, the convolution of the impulse response of a system and its inverse, if it exists, resolves to the Kronecker delta.

One of the qualities of a linear shift-invariant system is the duration of its impulse response. If the impulse response contains only finitely many non-zero entries, then this system is referred to as a finite impulse response, or FIR, system. On the other hand, if the impulse response has infinitely many non-zero components, then the system is referred to as an infinite impulse response, or IIR, system. If these non-zero entries are located in specific regions about the origin, then we can also define additional attributes of the system.

Definition 3.2.1. A linear shift-invariant system is causal, or recursively computable, if its impulse response is a one-sided sequence. ⧫

An important feature of the two-dimensional linear shift-invariant system is the ability to shear its impulse response in order to form a new system with similar characteristics. The first example we have seen of this operation is from Lim [46], who mentions it only in passing. This ability has no one-dimensional precedent. We will refer to the system created by shearing the impulse response of an original system as the augmented system. We will say that two systems are related if one system can be obtained by augmenting the other.

These shearing operations are not of much interest in themselves. However, they provide us with important theoretical tools. These shearing operations allow us to prove an extension of Jensen's Inequality, which is the main goal of this chapter.

Definition 3.2.2. Let T be a linear, shift-invariant system, and let $h(n_1,n_2)$ be its impulse response. There are four basic ways to shear the impulse response. We can make horizontal shears or vertical shears. We will denote the right-shear by $h_{\rightarrow}(n_1,n_2) = h(n_1-n_2,n_2)$, and the down-shear by $h_{\downarrow}(n_1,n_2) = h(n_1,n_2-n_1)$. Each of these operations has an inverse, $h_{\leftarrow}(n_1,n_2) = h(n_1+n_2,n_2)$ and $h_{\uparrow}(n_1,n_2) = h(n_1,n_2+n_1)$, for the right and the down shearing of the impulse response respectively. These four shear operations will be referred to as the basic shear operations, and any augmented system can be obtained by performing a sequence of these operations on the original system. ⧫

While the formal mathematical definition for these system augmentations is necessary, it is useful to picture these operations in a longer form. In the following notation, we begin with the original impulse response of a system on the left. The arrow operation tells us which of the four basic shear operations has been performed on the system, and the right-hand side is the impulse response of the augmented filter. Recall that entries with no value in the impulse response are assumed to be zero.

$$\begin{array}{cccc} & & \times & a_1 \\ & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \end{array} \rightarrow \begin{array}{cccc} \times & a_1 & & \\ a_2 & a_3 & a_4 & \\ a_5 & a_6 & a_7 & a_8 \end{array} \tag{3.3}$$

$$\begin{array}{cccc} & & \times & a_1 \\ & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ & & & \end{array} \downarrow \begin{array}{cccc} a_5 & a_2 & \times & \\ & a_6 & a_3 & a_1 \\ & & a_7 & a_4 \\ & & & a_8 \end{array} \tag{3.4}$$

Shown above are two examples of this operation being performed on a causal system with a finite impulse response; the × denotes the origin. The two operations shown are the right and down shear operations since these two are the most important of the basic shear operations for our purposes. Note that for the right-shear operation, all rows below the origin (in the positive direction for the n_2-axis) shear to the right by an amount equal to how many rows away they are from the origin. Any rows above the origin would shear to the left in a similar fashion. For the down-shear operation, we see the same effect on the columns instead of the rows. The column containing the origin stays stationary, while the columns to the right of the origin (in the postive direction for the n_1-axis) shear downwards by an amount corresponding to that column's distance from the origin. The columns to the left of the origin will shear upwards in a similar fashion.

It is important to note that these shear operations are not commutative. For example, if we first apply a right-shear, and then a down-shear, we end up with the following impulse response:

$$\begin{array}{cccc} & & \times & a_1 \\ & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \end{array} \rightarrow \begin{array}{cccc} \times & a_1 & & \\ a_2 & a_3 & a_4 & \\ a_5 & a_6 & a_7 & a_8 \end{array} \downarrow \begin{array}{cccc} \times & & & \\ a_2 & a_1 & & \\ a_5 & a_3 & & \\ & a_6 & a_4 & \\ & & a_7 & \\ & & & a_8 \end{array}$$

If we were to first apply a down-shear, and then a right-shear, we would end up with this impulse response:

$$\begin{array}{cccc} & & \times & a_1 \\ & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \end{array} \downarrow \begin{array}{cccc} a_5 & a_2 & \times & \\ & a_6 & a_3 & a_1 \\ & & a_7 & a_4 \\ & & & a_8 \end{array} \rightarrow \begin{array}{ccccccc} a_5 & a_2 & \times & & & & \\ & & a_6 & a_3 & a_1 & & \\ & & & & a_7 & a_4 & \\ & & & & & & a_8 \end{array}$$

The end result of these two sequences of operations are not equal. Therefore, we can conclude that the order in which we apply the basic shear operations is important. This does not imply that there is a unique set of shear operations to travel between related systems. Indeed there are infinitely many ways to get from one to the other. We are usually interested in the existence of a sequence of operations, rather than finding a specific sequence.

The next important property of a system is its stability. We say that a system is stable if for any bounded input, the ouput of the system is also bounded. In practical terms, this means that our system will give reasonable output as long as we kindly give it reasonable input. Since the impulse response entirely characterizes a linear shift-invariant system, it also contains all the information we need about the stability of the system.

Definition 3.2.3. A linear shift-invariant system is stable if and only if its impulse response is absolutely summable [46]. That is,

$$\sum_{n_1} \sum_{n_2} |h(n_1, n_2)| < \infty. \qquad \blacklozenge$$

From this definition it is easy to see that the augmented system is stable if and only if the original system is stable, since we are summing all the same information. In fact, we could rearrange the impulse response in any order we choose and still retain stability, so why are these particular shear operations important? There are two reasons why we devote attention to these shear operations, and both have to do with the inverse system.

The first reason is that tests for the stability of the inverse of a system typically require the impulse response to be of a quarter-plane design. That is to say the impulse response of a quarter-plane system has non-zero entries in only one quadrant of the plane. Such an impulse response might look like the following:

$$\begin{array}{ccc} \times & a_1 & a_2 \\ a_3 & a_4 & a_5 \\ a_6 & a_7 & a_8 \end{array}$$

where the × denotes the origin. Notice that there are no non-zero entries to the left of the column containing the origin, or above the row containing the origin. This is what is meant by an impulse response of quarter-plane shape. A system with an impulse response of quarter-plane shape will be referred to as a quarter-plane system. Any causal system with a finite impulse response can be augmented to a quarter-plane system by applying enough right-shears.

The second reason for studying system augmentations is that the shearing operation and the operation of taking an inverse are commutative.

Theorem 3.2.4. *Given a linear shift-invariant system T and its inverse T^{-1} together with their impulse responses $h(n_1, n_2)$ and $h^{-1}(n_1, n_2)$ respectively, and the augmented system given by one of the basic shear operations T_a, the impulse response of the inverse of T_a is equal to the same basic shear operation performed on the impulse response of T^{-1}.*

Proof. We must find $h_a^{-1}(\ell_1, \ell_2)$ such that

$$\sum_{\ell_1}\sum_{\ell_2} h_a^{-1}(\ell_1, \ell_2) h_a(n_1 - \ell_1, n_2 - \ell_2) = \delta(n_1, n_2).$$

where $\delta(n_1, n_2)$ is the Kronecker delta. We will show the proof for the right-shear operation, since proving the result for all other basic shear operations follows the same principle.

$$\begin{aligned}
\delta(n_1, n_2) &= \delta(n_1 - n_2, n_2) \\
&= \sum_{m_1}\sum_{m_2} h^{-1}(m_1, m_2) h(n_1 - n_2 - m_1, n_2 - m_2) \\
&= \sum_{\ell_1}\sum_{\ell_2} h^{-1}(\ell_1 - \ell_2, \ell_2) h(n_1 - n_2 - (\ell_1 - \ell_2), n_2 - \ell_2) \\
&= \sum_{\ell_1}\sum_{\ell_2} h^{-1}(\ell_1 - \ell_2, \ell_2) h(n_1 - \ell_1 - (n_2 - \ell_2), n_2 - \ell_2) \\
&= \sum_{\ell_1}\sum_{\ell_2} h^{-1}(\ell_1 - \ell_2, \ell_2) h_{\rightarrow}(n_1 - \ell_1, n_2 - \ell_2).
\end{aligned}$$

Thus $h^{-1}(\ell_1 - \ell_2, \ell_2)$ is equivalent to $h_{\rightarrow}^{-1}(\ell_1, \ell_2)$. As mentioned before, this proof works for the other basic shear operations. ■

Corollary 3.2.5. *Given a linear shift-invariant system T and its inverse T^{-1}, the inverse of any augmented system, T_a, related to T can be found by applying the same augmentation to T^{-1}. That is, $\left(T^{-1}\right)_a = (T_a)^{-1}$.*

Proof. Since any augmentation of a linear shift-invariant system can be achieved by applying the basic shear operations on the impulse response in succession, we can just apply the above theorem after each basic shear operation. The result follows. ■

As mentioned earlier, we can move the components of the impulse response of a stable system around in any way we please to obtain a system that is equivalently stable. However, it is important to note that we cannot move the components of the impulse response of a system around in any fashion and expect the resulting system to have a stable inverse.

Example 3.2.6. Consider a system with the following impulse response

$$1_{\times} \quad 0.6 \quad 0.6$$

where the 1 is at the origin. The 1-D theory covers this case, and we can determine that the inverse of this system is stable. However, the system with impulse response

$$\begin{array}{cc} 1_{\times} & 0.6 \\ 0.6 & \end{array}$$

is not an augmentation of the original system and does not have a stable inverse. This statement will be proven in detail after a discussion about the z-transform. ♣

3.3 Augmentations and Transfer Functions

In this section, we will introduce the 2-D z-transform, and use transfer functions to describe the characteristics of a 2-D system.

Definition 3.3.1. The z-transform of a sequence $x(n_1, n_2)$ is defined as

$$X(z_1, z_2) = \mathcal{Z}[x(n_1, n_2)](z_1, z_2) = \sum_{n_1} \sum_{n_2} x(n_1, n_2) z_1^{-n_1} z_2^{-n_2}. \qquad \blacklozenge$$

We will denote a sequence using lower case letters, and the z-transform of that sequence will use upper case letters. $\mathcal{Z}$ will denote the operation of taking a z-transform of a sequence.

Definition 3.3.2. Given two summable 2-D sequences, $x_1(n_1, n_2)$ and $x_2(n_1, n_2)$, their convolution is defined as

$$x_1(n_1, n_2) \overset{2}{*} x_2(n_1, n_2) = \sum_{m_1} \sum_{m_2} x_1(m_1, m_2) x_2(n_1 - m_1, n_2 - m_2). \qquad \blacklozenge$$

Theorem 3.3.3 (Convolution Theorem). *Given two summable 2-D sequences, $x_1(n_1, n_2)$ and $x_2(n_1, n_2)$, the z-transform of their convolution is given by*

$$\mathcal{Z}[x_1(n_1, n_2) \overset{2}{*} x_2(n_1, n_2)](z_1, z_2) = X_1(z_1, z_2) X_2(z_1, z_2).$$

This theorem tells us that the z-transform of a convolution of two sequences results in a multiplication of the z-transforms of each sequence. This is akin to the convolution theorem for the Fourier transform [61].

The transfer function of a system is given by the ratio of the transformed output sequence to the transformed input sequence. For a linear shift-invariant system, recall from Equation (3.2) that the output is related to the input by the convolution

$$y(n_1, n_2) = x(n_1, n_2) \overset{2}{*} h(n_1, n_2),$$

where $h(n_1, n_2)$ is the impulse response of the system. Taking the z-transform of both sides of this equation, and using Theorem (3.3.3) gives us

$$Y(z_1, z_2) = X(z_1, z_2)H(z_1, z_2).$$

Therefore, the transfer function of a linear shift-invariant system is given by the z-transform of its impulse response [46]. That is,

$$\frac{Y(z_1, z_2)}{X(z_1, z_2)} = H(z_1, z_2) = \sum_{n_1} \sum_{n_2} h(n_1, n_2) z_1^{-n_1} z_2^{-n_2}.$$

The purpose of the following is to show that the transfer functions of related systems are also of interest. This will culminate in an extension of Jensen's Inequality, which we will directly apply to our noise shaping theory.

Consider our basic right-shear of the system T with impulse response $h(n_1, n_2)$. The augmented system has the impulse response $h_{\rightarrow}(n_1, n_2) = h(n_1 - n_2, n_2)$. This implies that the transfer function of this augmented system is

$$\begin{aligned} H_{\rightarrow}(z_1, z_2) &= \sum_{n_1} \sum_{n_2} h(n_1, n_2) z_1^{-(n_1+n_2)} z_2^{-n_2} \\ &= \sum_{n_1} \sum_{n_2} h(n_1, n_2) z_1^{-n_1} (z_1 z_2)^{-n_2} \\ &= H(z_1, z_1 z_2). \end{aligned}$$

Similarly, the other three basic shear operations act on the transfer function as follows:

$$\begin{aligned} H_{\downarrow}(z_1, z_2) &= H(z_1 z_2, z_2), \\ H_{\leftarrow}(z_1, z_2) &= H(z_1, z_1^{-1} z_2), \\ H_{\uparrow}(z_1, z_2) &= H(z_1 z_2^{-1}, z_2). \end{aligned}$$

These operations can be composed together to get the transfer function for any related system.

Recall in the 1-D theory that a causal system with transfer function $H(z)$ is stable if and only if its poles are inside the unit circle in the z-plane. Similarly, a causal, stable system is minimum-phase if and only if its zeros are inside the unit circle in the z-plane as well. For 2-D causal systems, this characterization of minimum-phase no longer works since poles and zeros of $H(z_1, z_2)$ are no longer isolated. This has been a sticking point in a large portion of the literature, where authors insist on trying to define a system as minimum-phase based on the location of its poles and zeros [20, 57]. Instead we will go back to our original definition of minimum-phase and forget about the location of poles and zeros for the moment.

Definition 3.3.4. A system is called minimum-phase if it is causal and stable with a causal and stable inverse. ⧫

Note that the name "minimum-phase" for 1-D systems actually comes from the term "minimum phase lag", which describes a characteristic of the phase of such a system. A minimum-phase system has the least phase lag possible for a given log-magnitude system response. In 2-D, we are borrowing the name because of its common definition, and not because of any properties of the phase surfaces generated by these systems. We will look into the properties of the phase surface of these systems later on, and use the terminology without much concern for physcial interpretation at the moment.

Also, note that in 1-D we had the case of marginally minimum-phase systems, which have zeros located on the unit circle $|z| = 1$. We will run into 2-D systems that we will also refer to as marginally minimum-phase.

In order to test if a system is minimum-phase or not, we need to be able to determine the stability of the inverse of the system. The following is one of the earliest results regarding this question.

Theorem 3.3.5 (Shanks's Theorem [46]). *A quarter-plane system with the transfer function $H(z_1, z_2) = \frac{1}{B(z_1,z_2)}$, where B is a polynomial, is stable if and only if $B(z_1, z_2) \neq 0$ for any (z_1, z_2) pair with $|z_1|, |z_2| \geq 1$ simultaneously.*

For computational reasons, Shanks's theorem is rarely used, and other tests for stability are employed. One of the more popular tests is Huang's Theorem. This theorem is still computationally intensive, but is a vast improvement over the amount of work needed to verify Shanks's Theorem in practice.

Theorem 3.3.6 (Huang's Theorem [30]). *A quarter-plane system with the transfer function $H(z_1, z_2) = \frac{1}{B(z_1,z_2)}$, where B is a polynomial, is stable if and only if $B(z_1, z_2) \neq 0$ for $|z_1| = 1, |z_2| \geq 1$ and $B(z_1, z_2) \neq 0$ for $|z_1| \geq 1, z_2 = 1$. By symmetry, this result holds true if we switch the roles of z_1 and z_2.*

Huang's Theorem states that in order to test stability, we have to check the zeros of the transfer function when holding z_1 (or z_2) fixed at all values on the unit circle, and additionally check the zeros of the transfer function when holding z_2 (or z_1) fixed at only one value.

In both of these theorems, the inverse is considered to be marginally stable if we find a zero in the region $|z_1| = |z_2| = 1$, but not in the region $|z_1| = 1, |z_2| > 1$, or $|z_2| = 1, |z_1| > 1$.

What Huang's theorem tells us is that if we have a stable quarter-plane FIR system with transfer function $B(z_1, z_2)$, then its inverse is also a stable quarter-plane system if the zeros of $B(z_1, z_2)$ are located in the appropriate regions. Compare this with the 1-D case, where we can determine if a system is minimum-phase based on the location of the poles and zeros of its transfer function. Note that a quarter-plane system is also a causal system, although the converse is not necessarily true; there are many causal systems that are not quarter-plane. Causal systems that are not of quarter-plane shape necessarily have a zero in the region $|z_1| > 1, |z_2| > 1$. Therefore, tests such as Huang's theorem and Shanks's theorem do not apply to these cases directly.

Example 3.3.7. If we refer back to Example 3.2.6, we can see that this system is quarter-plane and the transfer function $1 + 0.6z_1^{-1} + 0.6z_2^{-1}$ has a zero at $z_1 = -1, z_2 = -1.5$, which violates Huang's theorem. Hence this system has an unstable inverse. ♣

Example 3.3.8. Consider a system whose transfer function is $B(z_1, z_2) = 1 - \frac{1}{2}z_1 z_2^{-1}$. The impulse response of this system looks like

$$\begin{array}{cc} & 1_\times \\ -\frac{1}{2} & \end{array}$$

where the 1 is situated at the origin. The zeros of $B(z_1, z_2)$ are located at $z_1 = 2z_2$. So, for instance, $z_1 = 4, z_2 = 2$ is a zero of this transfer function, which is outside the region imposed by Shanks's theorem. The impulse response of the inverse system looks like

$$\begin{array}{ccccc} & & & & 1_\times \\ & & & \frac{1}{2} & \\ & & \frac{1}{4} & & \\ & \frac{1}{8} & & & \\ \iddots & & & & \end{array}$$

Again, the 1 is situated at the origin. Notice that the impulse response in both of these cases is absolutely summable, and that they are both causal. So, by our definition, this system is minimum-phase. Since this is not a quarter-plane system, Huang's theorem and Shanks's theorem are not applicable, and the zero at $z_1 = 4, z_2 = 2$ does not give us any information about the stability of the inverse system.

However, if we apply a right-shear to this system, the transfer function becomes

$$1 - \frac{1}{2}z_2^{-1}.$$

This system has all of its zeros on the plane $z_2 = \frac{1}{2}$. Notice that this is now a quarter-plane system, and that it satisfies Shanks's Theorem. Therefore, this system is minimum-phase. Since this system is related to the original system, we deduce that the original system is causal and stable, and has a causal and stable inverse. Therefore it is also minimum-phase, which is in agreement with our statement above.

This is not the only augmentation we could have performed to obtain a related quarter-plane system. If we apply a second right-shear, the transfer function becomes

$$1 - \frac{1}{2}z_1^{-1}z_2^{-1}.$$

This is still a quarter-plane system, so Huang's theorem and Shanks's theorem still apply. For the zeros of this transfer function, we find that if $|z_1| \geq 1$ then $|z_2| \leq \frac{1}{2}$, and if $|z_2| \geq 1$ then $|z_1| \leq \frac{1}{2}$ by symmetry. So this related system also satisfies Shanks's Theorem.

♣

In essence, we can augment any FIR system to a quarter-plane system and use Shanks's Theorem and Huang's Theorem to test the stability of the related system in order to determine the stability properties of the original system. This will be the logic we use to determine if a system is minimum-phase, since in all but very simple cases, finding the impulse response of the inverse and checking its stability is infeasible.

In order to actually apply Huang's Theorem in practice, we use what is called a Root Map. This is a mapping of the zeros of the transfer function if we hold one of the values at a constant value of magnitude 1. Say we hold z_1 fixed. If we plot the zeros of a transfer function in the z_2 plane for every value of $|z_1| = 1$, we will obtain a continuous curve of zeros. If this curve of zeros should cross outside the unit circle of the z_2 plane, then we know that there exists a point (z_1, z_2) such that $|z_1| = 1$ and $|z_2| > 1$, and therefore this transfer function fails the criteria of Huang's Theorem. We will see root maps used in later chapters.

3.4 Extending Jensen's Inequality

The transfer function allows us to see how a system will behave if we input a sinusoidal signal of a certain frequency. The two most common tools to analyze a system are its log-magnitude and its phase. The log-magnitude indicates how much the system will amplify the input, and the phase will indicate how many degrees the output of the system leads or lags behind the input signal. For a discrete linear shift-invariant system with transfer function $H(z_1, z_2)$, we obtain the log-magnitude by evaluating $\log\left(\left|H(e^{i\theta_1}, e^{i\theta_2})\right|\right)$ for values of θ_1 and θ_2. The phase is found by evaluating the argument of the complex number $H(e^{i\theta_1}, e^{i\theta_2})$.

The problem we face is that different transfer functions can give the same log-magnitude, but different phases. It turns out that, for the application to noise shaping, given a log-magnitude shape we can find a system that produces this log-magnitude shape and has an optimal phase associated with it. We will build towards an algorithm to find such a system, but first we will prove that minimum-phase systems are the optimal systems to use in noise shaping.

To begin, we will show that the log-magnitude of related systems also share certain important characteristics.

Lemma 3.4.1. *Let $H(z_1, z_2)$ be the transfer function of a system. Let $H_a(z_1, z_2)$ be an augmentation of H. Then,*

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \log\left(\left|H\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1\ d\theta_2 = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \log\left(\left|H_a\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1\ d\theta_2.$$

Proof. We will prove this lemma for the basic right-shear. The proof for the other three basic shears follows the same methods.

Let $D_{\theta_1\theta_2}$ denote the square $[-\pi, \pi] \times [-\pi, \pi]$. See Figure 3.1(a). We have

$$\begin{aligned} & \iint_{D_{\theta_1\theta_2}} \log\left(\left|H_{\rightarrow}\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1\ d\theta_2 \\ = & \iint_{D_{\theta_1\theta_2}} \log\left(\left|H\left(e^{i\theta_1}, e^{i(\theta_1+\theta_2)}\right)\right|\right) d\theta_1\ d\theta_2. \end{aligned}$$

Let $\phi_1 = \theta_1$ and $\phi_2 = \theta_1 + \theta_2$. Then we have that the Jacobian is

$$\left|\frac{\partial(\theta_1, \theta_2)}{\partial(\phi_1, \phi_2)}\right| = \left|\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}\right| = 1.$$

The transformed domain $D_{\phi_1\phi_2}$ is a parallelogram with vertices at $(-\pi, -2\pi), (-\pi, 0), (\pi, 2\pi), (\pi, 0)$. See Figure 3.1(b).

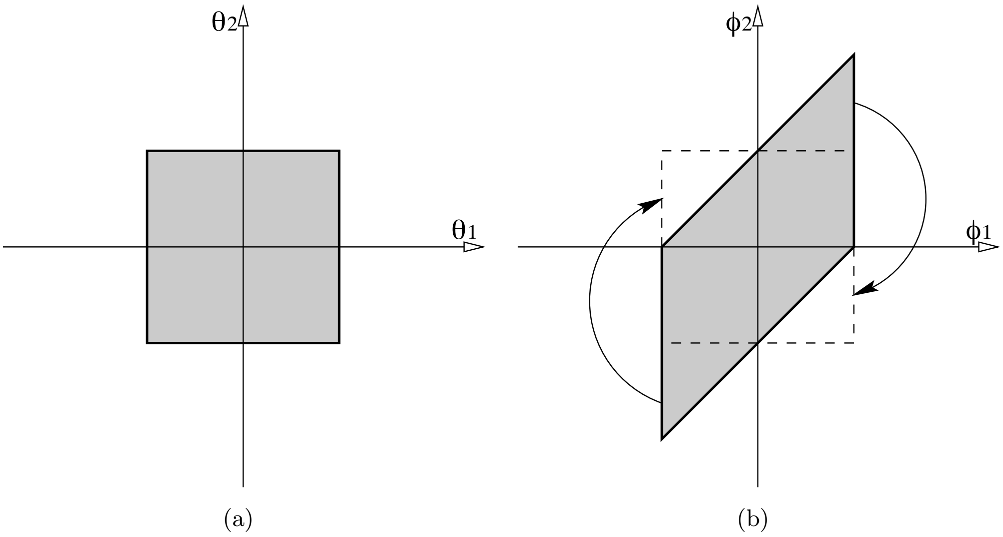


Figure 3.1: (a) $D_{\theta_1\theta_2}$. (b) $D_{\phi_1\phi_2}$, taking advantage of periodicity.

Thus we have

$$\begin{aligned} &\iint_{D_{\theta_1\theta_2}} \log\left(\left|H\left(e^{i\theta_1}, e^{i(\theta_1+\theta_2)}\right)\right|\right) d\theta_1\ d\theta_2 \\ = &\iint_{D_{\phi_1\phi_2}} \log\left(\left|H\left(e^{i\phi_1}, e^{i\phi_2}\right)\right|\right) d\phi_1\ d\phi_2. \end{aligned}$$

Since $H(e^{i\phi_1}, e^{i\phi_2})$ is $2\pi \times 2\pi$ periodic, integrating over the region $D_{\phi_1\phi_2}$ is equivalent to integrating over the square $[-\pi, \pi] \times [-\pi, \pi]$. See Figure 3.1(b). Thus our last integral is equivalent to

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \log\left(\left|H\left(e^{i\phi_1}, e^{i\phi_2}\right)\right|\right) d\phi_1\ d\phi_2,$$

which completes the proof for the case of the right-shear operation.

As mentioned earlier, this proof will work for the other three basic shear operations. Since any augmented system can be obtained via successive shear operations, we conclude that any two related systems of this form give the same value of this integral. ■

This lemma gives us another insight into the nature of related systems. We now know that the volume underneath the log-magnitude response surface of a system is invariant under the shear operation.

Before we can generalize Jensen's Inequality, there are a couple of technical details to get out of the way.

Theorem 3.4.2 (Rouché's Theorem [15]). *Supose $f(z)$ and $g(z)$ are analytic inside and on a simple closed contour C, and that $|f(z)| > |g(z)|$ at each point on C. Then $f(z)$ and $f(z)+g(z)$ have the same number of zeros inside C.*

Rouché's theorem is a standard result in complex analysis. This theorem is the basis for the proof of the following lemma, which was given by Marden.

Lemma 3.4.3 ([50]). *Let*

$$f(z) = a_0 + a_1 z + \cdots + a_n z^n = a_n \prod_{j=1}^{p} (z - z_j)^{m_j},$$

where $a_n \neq 0$ and $m_1 + \cdots + m_p = n$. Also, let

$$F(z) = (a_0 + \epsilon_0) + (a_1 + \epsilon_1)z + \cdots + (a_{n-1} + \epsilon_{n-1})z^{n-1} + a_n z^n,$$

and let $0 < r_k < \min |z_k - z_j|$ for $j = 1, \ldots, k-1, k+1, \ldots, p$. There exists a positive number ϵ such that if $|\epsilon_j| \leq \epsilon$ for each j, then $F(z)$ has precisely m_k zeros in the circle with centre z_k and radius r_k.

Proof. Consider the closed circle C_k with centre z_k and radius r_k, and the polynomial

$$g(z) = \epsilon_0 + \epsilon_1 z + \cdots + \epsilon_{n-1} z^{n-1}.$$

Let

$$N_k = \sum_{j=0}^{n-1} (r_k + |z_k|)^j .$$

Then $|g(z)| \le \epsilon N_k$. Note that for $z \in C_k$, we also know that

$$\begin{aligned} |f(z)| &= |a_n| \, r_k^{m_k} \prod_{j=1, j\neq k}^{p} |(z - z_j)|^{m_j} \\ &\ge |a_n| \, r_k^{m_k} \prod_{j=1, j\neq k}^{p} (|z_j - z_k| - r_k)^{m_j} \\ &= \Gamma_k > 0. \end{aligned}$$

Choose $\epsilon < \frac{\Gamma_k}{N_k}$, and we have $|g(z)| < |f(z)|$ on C_k. Hence, according to Rouché's theorem, we have that $F(z)$ has the same number of zeros in C_k as $f(z)$. By our choice of r_k, we know that the only zeros of $f(z)$ inside C_k is the zero z_k of multiplicity m_k. So there are m_k zeros of $F(z)$ inside C_k.

If we choose ϵ such that $\epsilon < \frac{\Gamma_j}{N_j}$ for $j = 1, \ldots, p$, then the conclusion holds for all circles, C_j, simultaneously. ■

Lemma 3.4.3 is a technical way of saying that the zeros of a polynomial move continuously as we vary the polynomial coefficients in a continuous manner. In particular, we are going to have a polynomial of the form $f(z) = a_0(y) + a_1(y)z + a_2(y)z^2 + \cdots + a_n(y)z^n$, and we will need to keep track of how the zeros of this polynomial are moving as we vary the parameter y in the continuous coeffecient functions $a_j(y)$.

We now have all the information we need to extend Jensen's inequality to 2-D.

Theorem 3.4.4 (Generalized Jensen's Inequality). *Given a causal and stable FIR system with transfer function* $1 - H(z_1, z_2)$*, where* $H(z_1, z_2)$ *has no constant term, we have*

$$\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log \left(\left| 1 - H\left(e^{i\theta_1}, e^{i\theta_2}\right) \right| \right) d\theta_1 d\theta_2 \ge 0$$

with equality if and only if $1 - H(z_1, z_2)$ *is minimum-phase or marginally minimum-phase.*

Proof. Instead of evaluating this integral for $1 - H(z_1, z_2)$, we will evauate it for a related system $1 - H_a(z_1, z_2)$. By Lemma 3.4.1, these two integrals will be the same.

Since $H(z_1, z_2)$ contains only finitely many terms, we can apply several basic right-shears to this system until each column in the augmented impulse response contains at most one entry. For example, if the system with transfer function $H(z_1, z_2)$ has an impulse response that looks like

$$\begin{array}{ccc} & \times & a_1 \\ a_3 & a_4 & a_5 \\ a_6 & a_7 & a_8 \end{array}$$

where the × denotes the origin, then applying three right-shear operations gives us

$$\begin{array}{ccc} & \times & a_1 \\ a_3 & a_4 & a_5 \\ a_6 & a_7 & a_8 \end{array} \xrightarrow{3} \begin{array}{cccccccc} \times & a_1 & & & & & & \\ & & a_3 & a_4 & a_5 & & & \\ & & & & & a_6 & a_7 & a_8 \end{array}$$

The transfer function, $H_{\xrightarrow{3}}(z_1, z_2)$, of this system can be written as $A_1(z_2)z_1^{-1} + A_2(z_2)z_1^{-2} + \cdots + A_8(z_2)z_1^{-8}$, where $A_j(z_2)$ is continuous for $z_2 \neq 0$. Now, if we apply one basic down-shear operation, then each row in the impulse response of this system contains only one entry as well. Now the transfer function, $H_{\xrightarrow{3}\downarrow}(z_1, z_2)$, can be written as $B_1(z_2)z_1^{-1} + B_2(z_2)z_1^{-2} + \cdots + B_8(z_2)z_1^{-8}$, where $B_j(z_2)$ is continuous for $z_2 \neq 0$, or equivalently as $C_1(z_1)z_2^{-1} + C_2(z_1)z_2^{-2} + \cdots + C_8(z_1)z_2^{-8}$, where $C_j(z_1)$ is also continuous for $z_1 \neq 0$. In particular, there is no constant term. Also, for any fixed value z_2, the transfer function reads as a causal transfer function in z_1, and vice versa.

In general, it will take n right-shear operations to create this system, and its transfer function will be denoted by $H_{\xrightarrow{n}\downarrow}(z_1, z_2)$. This is the transfer function we use in the integral expression.

For any fixed value of z_2, the inner integral of our expression is the same as in the one dimensional Jensen's theorem. For some value of z_2, where $|z_2| = 1$, suppose there exists a value of z_1 with $|z_1| = 1 + \alpha, \alpha > 0$ such that $1 - H_{\xrightarrow{n}\downarrow}(z_1, z_2) = 0$. Then, for the value of ϕ between $-\pi$ and π such that $e^{-i\phi} = z_2$, we have

$$\int_{-\pi}^{\pi} \log\left(\left|1 - H_{\xrightarrow{n}\downarrow}\left(e^{i\theta_1}, e^{i\phi}\right)\right|\right) d\theta_1 > 0,$$

and this transfer function is not minimum-phase.

Recall that we have augmented this system so that we can treat $1 - H_{\xrightarrow{n}\downarrow}(z_1, z_2)$ as a polynomial in z_1 with coefficients that depend continuously on z_2, with a leading 1 as the constant term. The leading 1 is important so that we can use Jensen's theorem. From Lemma 3.4.3, we know that the zeros move continuously along with our parameter z_2. Therefore, if there is a zero at (z_1, z_2), where $|z_1| = 1 + \alpha, |z_2| = 1$, then there exists an ϵ such that for any $\theta_2 \in (\phi - \epsilon, \phi + \epsilon)$, the expression $1 - H_{\xrightarrow{n}\downarrow}(z_1, e^{i\theta_2})$ has a zero for some value of z_1 where

$|z_1| > 1 + \frac{\alpha}{2}$. Thus

$$\int_{\phi-\epsilon}^{\phi+\epsilon} \int_{-\pi}^{\pi} \log\left(\left|1 - H_{\xrightarrow{n}\downarrow}\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1 \; d\theta_2 > 0.$$

If there is no such point z_2, where $|z_2| = 1$, then we must have that the integral expression evaluates to zero. This means that this transfer function has no zeros in the region $|z_2| = 1, |z_1| > 1$. Since we can also regard the transfer function as a polynomial in z_2 with coefficients that depend continuously on z_1, with a 1 as the leading constant term, this argument also applies if we switch the order of integration. That is, if the integral expression evaluates to zero, then we deduce that the transfer function cannot have any zeros in the region $|z_1| = 1, |z_2| > 1$. Therefore, by Huang's theorem, we conclude that $1 - H_{\xrightarrow{n}\downarrow}(z_1, z_2)$ is the transfer function of a minimum-phase system or a marginally minimum-phase system. Therefore, by Corollary 3.2.5, $1 - H(z_1, z_2)$ must be the transfer function of a minimum-phase system or a marginally minimum-phase system. ■

At first glance, limiting ourselves to causal systems with a transfer function of the form $1 - H(z_1, z_2)$ might seem like an unnecessary limitation, even though this class of transfer functions will cover all the cases we will need for the noise shaping theory. If we assume the constant term is some other non-zero value, then we can factor that value out of the transfer function. That is, $a - H(z_1, z_2) = a(1 - \frac{H(z_1,z_2)}{a})$, where a is a scalar. If we substitute this into our log-magnitude integral, we find that

$$\begin{aligned} & \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log\left(\left|a - H\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1 d\theta_2 \\ = & \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log(|a|) + \log\left(\left|1 - \frac{H\left(e^{i\theta_1}, e^{i\theta_2}\right)}{a}\right|\right) d\theta_1 d\theta_2 \\ \geq & \; 4\pi^2 \log(|a|). \end{aligned} \tag{3.5}$$

with equality if and only if the system with transfer function $1 - \frac{H(z_1,z_2)}{a}$ is minimum-phase. Note that multiplying the transfer function by a does not affect the stability of system or its inverse. Therefore, we can also say that equality holds in Equation (3.5) if and only if the system with transfer function $a - H(z_1, z_2)$ is minimum-phase.

If a is zero, then the system cannot be minimum-phase since the inverse system necessarily has to be acausal to correct for the overall delay. This matches up with the limit of the above case as $a \to 0$,

$$\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log\left(\left|-H\left(e^{i\theta_1}, e^{i\theta_2}\right)\right|\right) d\theta_1 d\theta_2 \geq -\infty,$$

which is not an insightful statement, but mathematically consistent.

It also needs to be mentioned that two related systems are, in fact, different systems. Each has its own log-magnitude and phase properties. We have shown that related systems share stability properties and volume under the log-magnitude surface. These are the only similarities between related systems with which we need to concern ourselves.

This extension of Jensen's Inequality also gives us an efficient test of whether a given FIR system is minimum-phase. Until now, the normal test has been to find the transfer function of the related quarter-plane system and find its zeros via a root map. This is a computationally intense process. With this extension of Jensen's Inequality, we no longer have to find a related quarter-plane system, and numerical integration procedures are typically much more computationally efficient than root finding methods. Hence this extension to Jensen's Inequality is not only of theoretical value, it provides us with a computationally attractive test for the minimum-phase property. Moreover, the inequality also allows us to quantify the instability of the inverse of a system.

3.5 Comments on the Phase of 2-D Minimum-Phase Systems

As mentioned earlier, we chose the name "minimum-phase" because of the common definition with the 1-D case, and not because of the actual properties of the phase of such a system. In 1-D, the term "minimum-phase" was derived from the expression "minimum phase". The minimum-phase system will have the least amount of phase of any system with a given log-magnitude.

Example 3.5.1. Consider the 1-D system with transfer function

$$1 + 2z^{-1} + 3z^{-2}$$

This system is not minimum-phase since it has zeros outside the unit circle in the z-plane. The minmium-phase system with the same log-magnitude has the transfer function

$$3 + 2z^{-1} + 1z^{-2}$$

Figure 3.2 shows the phase of these two systems. Note that the graph of the phase of the minimum-phase system is always above the graph for the non-minimum-phase system. This is true in general for minimum-phase systems in 1-D.

♣

Unfortunately, the same results are not true in 2-D. This next example will demonstrate.

Example 3.5.2. Consider the system with transfer function

$$1 - \frac{6}{16}z_1^{-1} - \frac{51}{16}z_1 z_2^{-1} + \frac{13}{16}z_2^{-1} - \frac{1}{16}z_1^{-1}z_2^{-1} + \frac{9}{16}z_1^2 z_2^{-2} + \frac{15}{16}z_1 z_2^{-2} + \frac{3}{16}z_2^{-2}$$

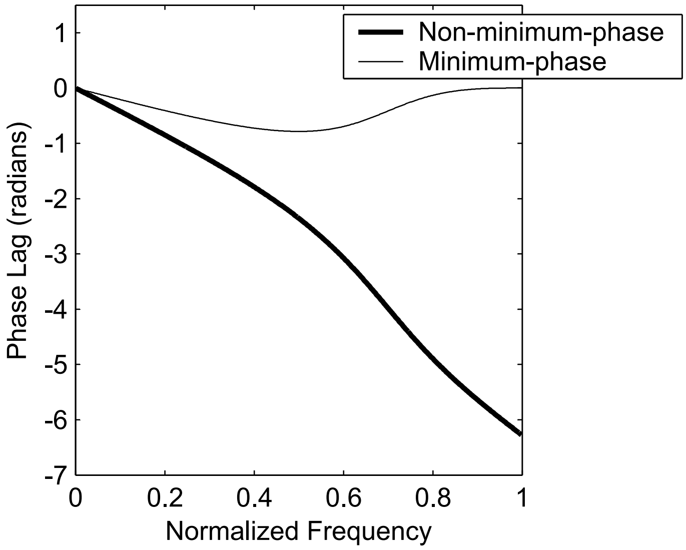


Figure 3.2: Example of 1-D phase.

This system is not minimum-phase. The phase of this system is shown in Figure 3.3.

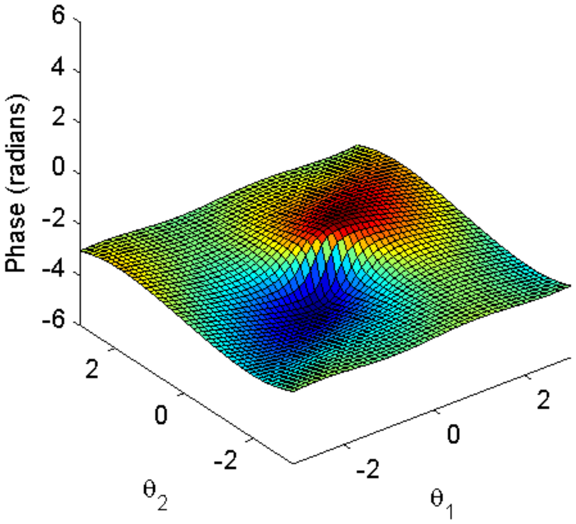


Figure 3.3: Phase of a 2-D non-minimum-phase system.

The minimum-phase system with the same log-magnitude has the transfer function

$$1 - \frac{6}{16}z_1^{-1} - \frac{25}{48}z_1z_2^{-1} - \frac{9}{48}z_2^{-1} - \frac{1}{16}z_1^{-1}z_2^{-1} + \frac{3}{48}z_1^2z_2^{-2} + \frac{5}{48}z_1z_2^{-2} + \frac{1}{48}z_2^{-2}$$

The phase of this system is shown in Figure 3.4.

The difference in phase between these two systems is shown in Figure 3.5.

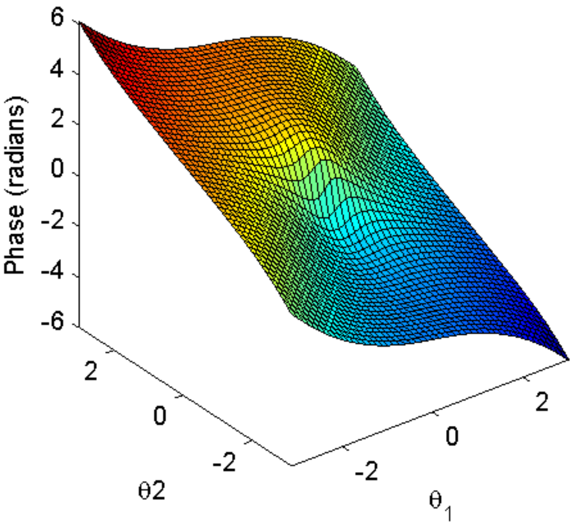
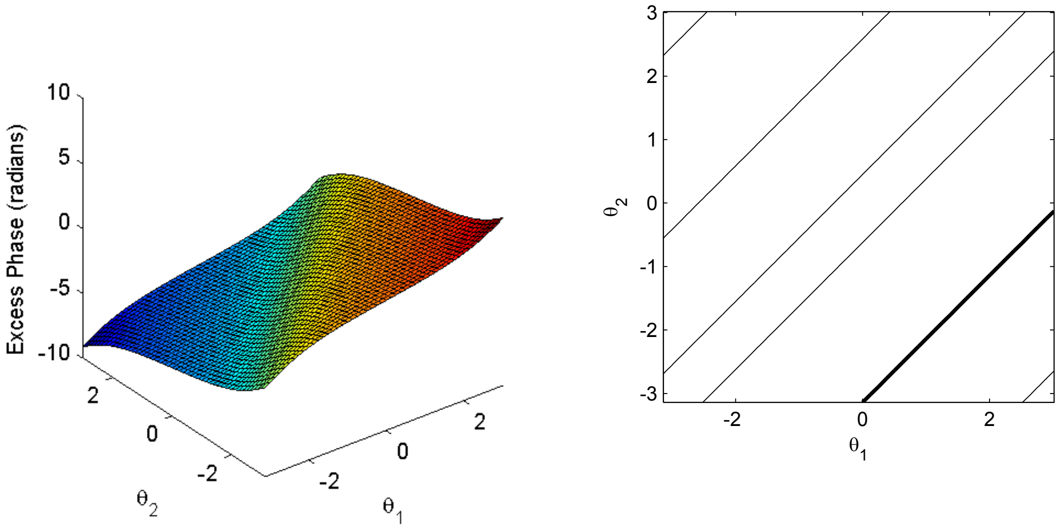


Figure 3.4: Phase of a 2-D minimum-phase system.



(a)

(b)

Figure 3.5: a) Excess phase. b) Contour plot of the excess phase. The thick line represents the zero-phase contour.

Notice that the difference in the phase is positive for some values of θ_1 and θ_2. This indicates that the non-minimum-phase system has less phase lag in this region than the "minimum-phase" system. ♣

As the above example shows, the term "minimum-phase" may be a misnomer in two dimensions.

3.6 Future Considerations

One of the remaining consideration is the uniqueness of a minimum-phase system for a given log-magnitude. It is believed that there is only one minimum-phase system for a given log-magnitude, but again we do not have a proof of this conjecture. In terms of the theory we are developing in this thesis, uniqueness of the minimum-phase system is not an issue. All we need is to be able to find one minimum-phase system with a given log-magnitude. This problem will be addressed in the next chapter.

Outside of dithered quantization and noise shaping, it would be interesting to see applications of this extension of Jensen's Inequality to other areas of the theory of functions of several complex variables.

It should be noted that the proofs in this chapter are all generalizable to higher dimensions.

Chapter 4

Design of a Noise Shaper

In previous chapters, we have explored the reasons that one would want to design a minimum-phase system. Several methods have been found in the literature[77, 35, 25, 57, 58, 19, 13] to accomplish this task, but each has counter-examples which show the method failing to achieve a minimum-phase system.

In this section, we will extend the Hilbert Transform method that was described in Section 1.6. We will prove that this method does, in fact, produce the impulse response of a minimum-phase system, and we will use this method to design a 2-D noise shaper that improves on the standard Floyd and Steinberg design for image processing.

4.1 The 2-D Discrete Fourier Transform

Before we can dig into the theory of a 2-D transform based on the Hilbert transform, we must first build some theory on the 2-D discrete Fourier transform and convolution.

Definition 4.1.1. The 2-D discrete Fourier transform of a 2-D finite sequence, $x(n_1, n_2)$, of size $N \times M$ is defined as

$$X(k_1, k_2) = \mathcal{F}_2\left[x(n_1, n_2)\right](k_1, k_2) = \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{M-1} x(n_1, n_2) e^{-i2\pi\left(\frac{k_1 n_1}{N} + \frac{k_2 n_2}{M}\right)}.$$

The inverse transform is defined as

$$x(n_1, n_2) = \mathcal{F}_2^{-1}\left[X(k_1, k_2)\right](n_1, n_2) = \frac{1}{NM}\sum_{k_1=0}^{N-1}\sum_{k_2=0}^{M-1} X(k_1, k_2) e^{i2\pi\left(\frac{k_1 n_1}{N} + \frac{k_2 n_2}{M}\right)}. \qquad \blacklozenge$$

We will often disregard the indices, and write the 2-D discrete Fourier transform as $\mathcal{F}_2\left\{x(n_1, n_2)\right\}$. In the future, we will end up writing many Fourier transforms in quick succession, and the omission of the transform indices will save considerable amounts of space, and lead to cleaner expressions.

DEFINITION 4.1.2. The partial discrete Fourier transforms of a 2-D finite sequence, $x(n_1, n_2)$, are defined as

$$x_{\leftrightarrow}(k_1, n_2) = \mathcal{F}_{\leftrightarrow}\left[x(n_1, n_2)\right](k_1, n_2) = \sum_{n_1=0}^{N-1} x(n_1, n_2) e^{-i2\pi\frac{k_1 n_1}{N}}$$

and

$$x_{\updownarrow}(n_1, k_2) = \mathcal{F}_{\updownarrow}\left[x(n_1, n_2)\right](n_1, k_2) = \sum_{n_2=0}^{M-1} x(n_1, n_2) e^{-i2\pi\frac{k_2 n_2}{M}}.$$

Their inverses are

$$x(n_1, n_2) = \mathcal{F}_{\leftrightarrow}^{-1}\left[x_{\leftrightarrow}(k_1, n_2)\right](n_1, n_2) = \frac{1}{N}\sum_{k_1=0}^{N-1} x_{\leftrightarrow}(k_1, n_2) e^{i2\pi\frac{k_1 n_1}{N}}$$

and

$$x(n_1, n_2) = \mathcal{F}_{\updownarrow}^{-1}\left[x_{\updownarrow}(n_1, k_2)\right](n_1, n_2) = \frac{1}{M}\sum_{k_2=0}^{M-1} x_{\updownarrow}(n_1, k_2) e^{i2\pi\frac{k_2 n_2}{M}}.$$

♦

Again, to save space, we will often write these transforms without the transform variables in order to save space. For example, we will often write $\mathcal{F}_{\leftrightarrow}\{x(n_1, n_2)\}$ for the partial transform across the rows of $x(n_1, n_2)$. These partial Fourier transforms are equivalent to taking a 1-D discrete Fourier transform across the rows or down the columns of $x(n_1, n_2)$ respectively.

THEOREM 4.1.3. *Given a 2-D finite sequence, $x(n_1, n_2)$, we have*

$$\mathcal{F}_2\{x(n_1, n_2)\} = \mathcal{F}_{\updownarrow}\{\mathcal{F}_{\leftrightarrow}\{x(n_1, n_2)\}\} = \mathcal{F}_{\leftrightarrow}\left\{\mathcal{F}_{\updownarrow}\{x(n_1, n_2)\}\right\}.$$

Proof. The proof is immediate from the definition of the 2-D discrete Fourier transform, and the fact that we can change the order of the summation. The proof for the inverse transforms is the same. ■

The consequence of this theorem is that we can break a 2-D discrete Fourier transform into two separate 1-D discrete Fourier transforms. This will become important in the upcoming sections.

DEFINITION 4.1.4. Given two 2-D finite sequences, $x(n_1, n_2)$ and $y(n_1, n_2)$, of size $N \times M$, the convolution of the two sequences is

$$x(n_1, n_2) \overset{2}{*} y(n_1, n_2) = \sum_{m_1=0}^{N-1}\sum_{m_2=0}^{M-1} x(m_1, m_2) y(n_1 - m_1, n_2 - m_2).$$

♦

Note that convolution is a commutative operation. That is, $x(n_1, n_2) \overset{2}{*} y(n_1, n_2) = y(n_1, n_2) \overset{2}{*} x(n_1, n_2)$.

Definition 4.1.5. Given two 2-D finite sequences, $x(n_1, n_2)$ and $y(n_1, n_2)$, of size $N \times M$, the partial convolutions of the two sequences are

$$x(n_1, n_2) *_{\leftrightarrow} y(n_1, n_2) = \sum_{m_1=0}^{M-1} x(m_1, n_2) y(n_1 - m_1, n_2)$$

and

$$x(n_1, n_2) *_{\updownarrow} y(n_1, n_2) = \sum_{m_2=0}^{N-1} x(n_1, m_2) y(n_1, n_2 - m_2).$$

♦

Theorem 4.1.6 (Convolution Theorem). *Given two sequences $x(n_1, n_2)$ and $y(n_1, n_2)$, the discrete Fourier transform of their pointwise product is*

$$\mathcal{F}_2 \{x(n_1, n_2) y(n_1, n_2)\} = \frac{1}{NM} \mathcal{F}_2 \{x(n_1, n_2)\} \overset{2}{*} \mathcal{F}_2 \{y(n_1, n_2)\}.$$

Similarly, given two sequences $X(k_1, k_2)$ and $Y(k_1, k_2)$, the inverse discrete Fourier transform of their pointwise product is

$$\mathcal{F}_2^{-1} \{X(k_1, k_2) Y(k_1, k_2)\} = \mathcal{F}_2^{-1} \{X(k_1, k_2)\} \overset{2}{*} \mathcal{F}_2^{-1} \{Y(k_1, k_2)\}.$$

Theorem 4.1.7. *Given two 2-D finite length sequences, $x(n_1, n_2)$ and $y(n_1, n_2)$, the Fourier transform of the pointwise product of these sequences is*

$$\begin{aligned} \mathcal{F}_2 \{x(n_1, n_2) y(n_1, n_2)\} &= \mathcal{F}_{\updownarrow} \Big\{ \tfrac{1}{N} \mathcal{F}_{\leftrightarrow} \{x(n_1, n_2)\} *_{\leftrightarrow} \mathcal{F}_{\leftrightarrow} \{y(n_1, n_2)\} \Big\} \\ &= \mathcal{F}_{\leftrightarrow} \Big\{ \tfrac{1}{M} \mathcal{F}_{\updownarrow} \{x(n_1, n_2)\} *_{\updownarrow} \mathcal{F}_{\updownarrow} \{y(n_1, n_2)\} \Big\}. \end{aligned}$$

Similarly, given two 2-D finite sequences, $X(k_1, k_2)$ and $Y(k_1, k_2)$, the inverse Fourier transform of the pointwise product is

$$\begin{aligned} \mathcal{F}_2^{-1} \{X(k_1, k_2) Y(k_1, k_2)\} &= \mathcal{F}_{\updownarrow}^{-1} \Big\{ \mathcal{F}_{\leftrightarrow}^{-1} \{X(k_1, k_2)\} *_{\leftrightarrow} \mathcal{F}_{\leftrightarrow}^{-1} \{Y(k_1, k_2)\} \Big\} \\ &= \mathcal{F}_{\leftrightarrow}^{-1} \Big\{ \mathcal{F}_{\updownarrow}^{-1} \{X(k_1, k_2)\} *_{\updownarrow} \mathcal{F}_{\updownarrow}^{-1} \{Y(k_1, k_2)\} \Big\}. \end{aligned}$$

Proof. This theorem is a consequence of Theorem 4.1.3, and the 1-D theory of Fourier transforms. ■

4.2 Extended Hilbert Transform Method

As mentioned earlier, we have not been able to find a relationship between the log-magnitude and phase of a 2-D minimum-phase system in the literature. In this chapter, we will prove that such a relationship does exist. Specifically, we begin with the notion that the inverse discrete Fourier transform of the sampled log-magnitude and phase is a one-sided sequence. From this intuitive beginning, we will prove that the Hilbert transform method can be extended to produce the impulse response of a 2-D minimum-phase system.

For simplicity and smaller notation, we will assume that all of our finite sequences are of size $N \times N$, where N is even. All of the arguments that follow are easily modified for the cases where we are using finite sequences of size $N \times M$ where N and M are different even numbers. Similar results will hold if we drop the restriction that N and M need to be even, but the proofs would need to be modified.

First, we need to define a notion of one-sidedness for 2-D finite length sequences. Our definition of a one-sided finite length sequence will mirror Definition 3.1.2, Marzetta's definition of causality. Recall that a 2-D one-sided sequence has the form

$$\begin{array}{ccccccc} & \vdots & & & & \vdots & \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & a_{0,0}\times & a_{1,0} & a_{2,0} & \cdots \\ \cdots & a_{-2,1} & a_{-1,1} & a_{0,1} & a_{1,1} & a_{2,1} & \cdots \\ \cdots & a_{-2,2} & a_{-1,2} & a_{0,2} & a_{1,2} & a_{2,2} & \cdots \\ & \vdots & & & & \vdots & \end{array} \tag{4.1}$$

Definition 4.2.1. Let $x(n_1, n_2)$ be a 2-D finite length sequence of size $N \times N$, N even. $x(n_1, n_2)$ is defined to be one-sided, or causal, if $x(n_1, 0) = 0$ for $n_1 = \frac{N}{2} + 1, \ldots, N-1$, and $x(n_1, n_2) = 0$ for $n_2 = \frac{N}{2} + 1, \ldots, N-1$. ♦

A finite length 2-D, one-sided sequence is just the first $\frac{N}{2}$ points away from the origin. These points need to be repeated periodically so that the origin is in the upper left position (this is done for consistency with our definition of the 2-D discrete Fourier transform). So if we take $N = 4$ and reference Expression (4.1), a finite length 2-D one-sided sequence is

$$\begin{array}{cccc} a_{0,0}\times & a_{1,0} & a_{2,0} & 0 \\ a_{0,1} & a_{1,1} & a_{2,1} & a_{-1,1} \\ a_{0,2} & a_{1,2} & a_{2,2} & a_{-1,2} \\ 0 & 0 & 0 & 0 \end{array}$$

Definition 4.2.2. Let $x(n_1, n_2)$ be a 2-D finite length sequence. The even part of $x(n_1, n_2)$ is defined to be

$$x_e(n_1, n_2) = \frac{x(n_1, n_2) + x(N - n_1, N - n_2)}{2},$$

where $x(N, n_2)$ is defined to be $x(0, n_2)$, and $x(n_1, N)$ is similarly defined to be $x(n_1, 0)$. The odd part of $x(n_1, n_2)$ is defined to be

$$x_o(n_1, n_2) = \frac{x(n_1, n_2) - x(N - n_1, N - n_2)}{2}.$$ ♦

Given these two definitions, we note that the even and odd parts of a one-sided 2-D finite length sequence are uniquely related by

$$x(n_1, n_2) = x_e(n_1, n_2) + sgn(n_1, n_2)x_e(n_1, n_2),$$

where

$$sgn(n_1, n_2) = \begin{cases} 1, & n_2 = 0, n_1 = 1, \ldots \frac{N}{2} - 1 \text{ or } n_2 = 1, \ldots, \frac{N}{2} - 1 \\ -1, & n_2 = 0, n_1 = \frac{N}{2} + 1, \ldots, N - 1 \text{ or } n_2 = \frac{N}{2} + 1, \ldots, N - 1 \\ 0, & \text{otherwise} \end{cases} .$$

Definition 4.2.3. Given a 2-D finite sequence C, we define the causalization transformation $\mathcal{K}_C$ as

$$\mathcal{K}_C \left[x(n_1, n_2)\right](m_1, m_2) = \mathcal{F}_2^{-1}\Big[\exp\big(\mathcal{F}_2\left[\mathcal{F}_2^{-1}\left[x(n_1, n_2)\right](k_1, k_2) \cdot C(k_1, k_2)\right]\big)\Big](m_1, m_2),$$

where the multiplication with the sequence C is component-wise (C and x must have the same dimensions). Compare this definition against the 1-D Hilbert transform method described in Section 1.6. ♦

We will prove that if we take the $N \times N$ finite sequence

$$C(n_1, n_2) = \begin{array}{cccccccc} 1_{\times} & 2 & \cdots & 2 & 1 & 0 & \cdots & 0 \\ 2 & 2 & \cdots & 2 & 2 & 2 & \cdots & 2 \\ \vdots & \ddots & & & & & & \vdots \\ 2 & 2 & \cdots & 2 & 2 & 2 & \cdots & 2 \\ 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{array},$$

and if $x(n_1, n_2)$ is a real and even 2-D finite sequence, then the causalization transform applied to $x(n_1, n_2)$ will approximate the impulse response of a minimum-phase system whose log-magnitude (sampled) is $x(n_1, n_2)$. Note that $C(n_1, n_2) = 1 + sgn(n_1, n_2)$. However, there are a lot of things going on in this transform simultaneously, and it is hard to prove directly that this transform will produce a minimum-phase impulse response. Hence, we will break the proof into a series of lemmas. Our goal will be to show that the output of this transform method converges to a one-sided real sequence as $N \to \infty$, where the transform size is $N \times N$. Once we have proven this much, we can use the Extended Jensen's Inequality to deduce that this sequence must be converging to the impulse response of a minimum-phase system.

The first thing we will do is break the causalizer $C(n_1, n_2)$ into several pieces. Let

$$C_1(n_2, n_2) = \begin{array}{cccccccc} 1_\times & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 2 & 2 & \cdots & 2 & 2 & 2 & \cdots & 2 \\ \vdots & \ddots & & & & & & \vdots \\ 2 & 2 & \cdots & 2 & 2 & 2 & \cdots & 2 \\ 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{array}$$

Next, let

$$C_2(n_1, n_2) = \begin{array}{cccccccc} -1_\times & -1 & \cdots & -1 & -1 & -1 & \cdots & -1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{array}$$

Finally, let

$$C_3(n_1, n_2) = \begin{array}{cccccccc} 1_\times & 2 & \cdots & 2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{array}$$

Then we find that $C(n_1, n_2) = C_1(n_1, n_2) + C_2(n_1, n_2) + C_3(n_1, n_2)$.

Lemma 4.2.4. *Given two finite 2-D sequences, A_1 and A_2, of the same size, we have*

$$\mathcal{K}_{A_1+A_2}\left[x(n_1,n_2)\right](m_1,m_2) = \mathcal{K}_{A_1}\left[x(n_1,n_2)\right](m_1,m_2) \overset{2}{*} \mathcal{K}_{A_2}\left[x(n_1,n_2)\right](m_1,m_2).$$

Proof.

$$\begin{aligned}
&\mathcal{F}_2^{-1}\left\{\exp\left(\mathcal{F}_2\left\{\mathcal{F}_2^{-1}\left\{x(n_1,n_2)\right\}\cdot(A_1+A_2)\right\}\right)\right\}\\
&= \mathcal{F}_2^{-1}\left\{\exp\left(\mathcal{F}_2\left\{\mathcal{F}_2^{-1}\left\{x(n_1,n_2)\right\}\cdot A_1\right\}\right)\exp\left(\mathcal{F}_2\left\{\mathcal{F}_2^{-1}\left\{x(n_1,n_2)\right\}\cdot A_2\right\}\right)\right\}\\
&= \mathcal{K}_{A_1}\left[x(n_1,n_2)\right](m_1,m_2) \overset{2}{*} \mathcal{K}_{A_2}\left[x(n_1,n_2)\right](m_1,m_2),
\end{aligned}$$

where the first equality comes from the linearity of the Fourier transform, and the second equality comes from the Convolution Theorem (Theorem 4.1.6). ■

Our strategy will be to break the causalization transform into the three pieces C_1, C_2 and C_3. After explaining what each piece does individually, we will put all the pieces back together using a 2-D convolution.

Lemma 4.2.5.

$$\mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_1}\left\{x(n_1,n_2)\right\}\right\}$$

produces a 2-D sequence whose columns converge to a 1-D one-sided sequence at a rate of $O(\frac{1}{N^2})$, and the first entry in each column is e^{a_i}, where a_i is the average value of the i^{th} column.

Proof. We are essentially doing the $\mathcal{K}$ transform without performing the last Fourier transform across the rows. Consider

$$\mathcal{F}_2\left\{\mathcal{F}_2^{-1}\left\{x(n_1,n_2)\right\}\cdot C_1\right\} = x(n_1,n_2) \overset{2}{*} \frac{1}{N^2}\mathcal{F}_2\left\{C_1\right\}$$

The Fourier transform of C_1 is equivalent to the 1-D Hilbert transform down the first column multiplied by a factor of N and zero elsewhere. Doing a 2-D convolution with this is equivalent to convolving each column individually with the one non-zero column of $\frac{1}{N^2}\mathcal{F}_2\left\{C_1\right\}$. Then

$$\mathcal{F}_{\updownarrow}^{-1}\left\{exp\left(x(n_1,n_2) \overset{2}{*} \frac{1}{N^2}\mathcal{F}_2\left\{C_1\right\}\right)\right\}$$

is equivalent to the 1-D Hilbert transform method on each of the columns. Therefore, the first entry in each column is exactly e^{a_i} and each column converges pointwise to a 1-D one-sided sequence at a rate of $O(\frac{1}{N^2})$ as a result of the 1-D theory. ■

Note that each column of $x(n_1, n_2)$ will not be an even sequence in general. This means that the result at the end of this lemma will have complex entries. However, for the $\mathcal{K}$ transform to produce a physically realizable system, all the entries of the resultant sequence have to be real. We will deal with this problem shortly, but first we will turn our attention to C_2.

Lemma 4.2.6.

$$\mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_2}\left\{x(n_1,n_2)\right\}\right\}$$

produces a 2-D sequence whose only entries are e^{-a_i} in the first entry of each column, where a_i is the average value of the i^{th} column.

Proof. As before, we are essentially ignoring the last Fourier transform across the rows. Consider

$$\mathcal{F}_2\left\{\mathcal{F}_2^{-1}\left\{x(n_1,n_2)\right\}\cdot C_2\right\} = x(n_1,n_2) \overset{2}{*} \frac{1}{N^2}\mathcal{F}_2\left\{C_2\right\}.$$

The Fourier transform of C_2 is a sequence whose first column is constant at $-N$, and all other entries are zero. Therefore, when convolving with $x(n_1,n_2)$, we obtain columns whose entries are all $-a_i$, where a_i is the average value of that column. Then we exponentiate each entry, and perform an inverse Fourier transform down each of the columns. We are left with exactly e^{-a_i} in the first entry of each column and zero elsewhere. ■

Now that we have these two pieces, we can combine them to tell us something meaningful.

Lemma 4.2.7.

$$\mathcal{K}_{C_1+C_2}\left\{x(n_1,n_2)\right\}$$

converges to a 2-D one-sided sequence, and whose only entry in the first row is a 1 in the first position.

Proof. Using Lemma 4.2.5 and Lemma 4.2.6, we obtain that

$$\begin{aligned}&\mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_1}\left\{x(n_1,n_2)\right\} \overset{2}{*} \mathcal{K}_{C_2}\left\{x(n_1,n_2)\right\}\right\}\\ &= \mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_1}\left\{x(n_1,n_2)\right\}\right\} *_{\updownarrow} \mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_2}\left\{x(n_1,n_2)\right\}\right\}.\end{aligned}$$

So without the last inverse Fourier transform across the rows, we are left with a column convolution of the pieces left behind by the last two lemmas. The column convolution consists of multiplying each entry in the i^{th} column of

$$\mathcal{F}_{\leftrightarrow}\left\{\mathcal{K}_{C_1}\left\{x(n_1,n_2)\right\}\right\}$$

by e^{-a_i}. Therefore, the first entry in each column is exactly 1, and each column is converging to a 1-D one-sided sequence at a rate of $O(\frac{1}{N^2})$. If we apply the final inverse Fourier transform across the rows of this resulting matrix, then the first row, containing all ones, becomes a one in the first entry, and zero elsewhere. The rows converging to zero continue to converge, and so the resulting 2-D finite sequence is still converging to a 1-D one-sided sequence down each column. ■

So far we have been able to prove that the first two pieces of our causalizer produce a causal output. Now, let's take a look at the remaining piece of the causalizer, C_3.

Lemma 4.2.8.

$$\mathcal{K}_{C_3}\left\{x(n_1, n_2)\right\}$$

produces a 2-D sequence whose first row converges to a 1-D one-sided sequence, and all other rows are zero. Additionally, the first entry in the first row is e^a*, where* a *is the average value of* $x(n_1, n_2)$*.*

Proof. Begin by applying the inverse Fourier transform to $x(n_1, n_2)$, and then multiplying component-wise by C_3. Now, apply the Fourier transform across the rows

$$\mathcal{F}_{\leftrightarrow}\left\{\mathcal{F}_2^{-1}\left\{x(n_1, n_2) \cdot C_3\right\}\right\}$$

Note that only the first row at this step is non-zero, since C_3 killed off all the other rows. For reference, we will call this particular sequence $x'(n_1, n_2)$. Next, apply the Fourier transform down each column, thus producing constant columns. Exponentiate the result and perform the inverse Fourier transform down each of the columns. The result is the same as if we just exponentiated the first row $x'(n_1, n_2)$, and left all other entries at zero. Finally, if we apply the inverse Fourier transform across each row, we find that the result in the first row is equivalent to applying the 1-D transform method to the first row of

$$\mathcal{F}_{\updownarrow}^{-1}\left\{x(n_1, n_2)\right\} \tag{4.2}$$

The first row of this 2-D sequence contains the average value of each of the columns of $x(n_1, n_2)$. Therefore, by the 1-D theory, the first row of the resulting sequence converges to a 1-D one-sided sequence, and the first entry in this row is e^a, where a is the average value of $x(n_1, n_2)$. ■

Corollary 4.2.9. *If* $x(n_1, n_2)$ *is a real and 2-D even sequence, then*

$$\mathcal{K}_{C_3}\left\{x(n_1, n_2)\right\}$$

produces a 2-D sequence whose first row converges to a 1-D one-sided, real sequence, and all other rows are zero. Additionally, the first entry in the first row is e^a *where* a *is the average value of* $x(n_1, n_2)$*.*

Proof. If $x(n_1, n_2)$ is a 2-D real, even sequence, then the first row of Expression (4.2) is real and even. This is because the i^{th} column and the $(N-i)^{\text{th}}$ column of $x(n_1, n_2)$ contain all the same entries, but in a different order (as required by the even-ness of the sequence). Hence the averages of these rows must be the same, and therefore, the first row of Expression (4.2) is real and 1-D even. The result follows by the 1-D theory. ■

Now that we have all the individual pieces, we can start to stick them all together.

Lemma 4.2.10.

$$\mathcal{K}_C\left\{x(n_1, n_2)\right\}$$

converges to a 2-D one-sided sequence at a rate of $O(\frac{1}{N^2})$, whose first entry in the first row is e^a, where a is the average value of $x(n_1, n_2)$.

Proof. Using Lemma 4.2.4, we know that

$$\mathcal{K}_C \{x(n_1, n_2)\} = \mathcal{K}_{C_1+C_2} \{x(n_1, n_2)\} \overset{2}{*} \mathcal{K}_{C_3} \{x(n_1, n_2)\}$$

Recall from Lemma 4.2.8 that the first row of $\mathcal{K}_{C_3} \{x(n_1, n_2)\}$ converges to a 1-D one-sided sequence and all other entries are zero. Therefore, the 2-D convolution is just a convolution of each row of $\mathcal{K}_{C_1+C_2} \{x(n_1, n_2)\}$ with the non-zero row of $\mathcal{K}_{C_3} \{x(n_1, n_2)\}$. Also, recall from Lemma 4.2.7 that the first row of $\mathcal{K}_{C_1+C_2} \{x(n_1, n_2)\}$ contains a one in the first entry and zero elsewhere. Therefore, the convolution with this row gives us the first row of $\mathcal{K}_{C_3} \{x(n_1, n_2)\}$ back again. The convolution with the remaining non-zero rows of $\mathcal{K}_{C_1+C_2} \{x(n_1, n_2)\}$ will give something non-zero, and the convolution with the rows converging to zero will continue to converge. Hence the resulting 2-D sequence will converge to a one-sided sequence. Furthermore, the first entry in the first row of this sequence is e^a, where a is the average value of $x(n_1, n_2)$. ■

We have been able to show that our transformation produces a 2-D causal sequence. This is a step in the right direction, but for this information to be useful, we also require that this transformation produce real output. If the input sequence, $x(n_1, n_2)$ is 2-D even, then it can be shown that the output will be real, as follows.

Lemma 4.2.11. *Consider the Hilbert transform method applied to a 1-D signal, $x(n)$, of length N, $\mathcal{H}\{x(n)\}$. If applying the 1-D transform to $x(n)$ gives an impulse response of a system whose transfer function is $e^a - H(z)$, then $\mathcal{H}\{x(N-n)\}$ gives the impulse response of a system whose transfer function is $e^a - \bar{H}(z)$, where $\bar{H}(z)$ is equivalent to $H(z)$ with all its coefficients conjugated.*

Proof. We know that the 1-D transform gives us a minimum-phase impulse response for a given real input $x(n)$. Therefore, $e^a - H(z)$ is the transfer function of a minimum-phase system. Hence, we can deduce that $e^a - \bar{H}(z)$ is also the transfer function of a minimum-phase system since conjugating all the coefficients of a polynomial is equivalent to conjugating all of its zeros.

Also, we find that by conjugating the zeros of the transfer function, we obtain a log-magnitude that is reflected about $\theta = 0$. Therefore, the log-magnitude of the system whose transfer function is $e^a - \bar{H}(z)$ (sampled) is equivalent to $x(N-n)$. ■

Lemma 4.2.12. *Let $x(n_1, n_2)$ be a real, 2-D even finite sequence. Then*

$$\mathcal{K}_{C_1+C_2} \{x(n_1, n_2)\}$$

produces a real 2-D finite sequence.

Proof. Consider $\mathcal{K}_{C_1}\{x(n_1, n_2)\}$. Since $x(n_1, n_2)$ is even, the i^{th} column of this 2-D sequence has the property that $x(i, n_2) = x(N - i, N - n_2)$. Therefore, if we're applying the 1-D transform to each column, then Lemma 4.2.11 applies. From the proof of Lemma 4.2.5, we know that if we don't perform the last inverse Fourier transform across the rows, then we have exactly done just the 1-D transform down each of the columns. Therefore, by Lemma 4.2.11, we find that without the last inverse Fourier transform across the rows, the real part of each row is an even sequence, and the imaginary part of each row is an odd sequence. The inverse Fourier transform of the real part stays real and even, and the transform of the imaginary part becomes real and odd, and the sum of the two is therefore real. Hence, every entry in $\mathcal{K}_{C_1}\{x(n_1, n_2)\}$ is real.

From Lemma 4.2.6, we know that $\mathcal{F}_{\leftrightarrow}\{\mathcal{K}_{C_2}\{x(n_1, n_2)\}\}$ contains a single row of non-zero entries, and those entries are in the first position of each column, with the value of e^{-a_i}. Since $x(n_1, n_2)$ is even, we know that $a_i = a_{N-i}$. Therefore, this row is an even, real sequence. Hence, performing the last inverse Fourier transform across the rows of $\mathcal{F}_{\leftrightarrow}\{\mathcal{K}_{C_2}\{x(n_1, n_2)\}\}$ gives a real, even sequence. The rest of the rows remain zero.

Now, $\mathcal{K}_{C_1+C_2}\{x(n_1, n_2)\}$, is the convolution of two real 2-D sequences by Lemma 4.2.4. This convolution has to be real. Therefore $\mathcal{K}_{C_1+C_2}\{x(n_1, n_2)\}$ is a real 2-D finite sequence. ■

Theorem 4.2.13. *Suppose $x(n_1, n_2)$ is a real, even 2-D finite sequence. Then*

$$\mathcal{K}_C\{x(n_1, n_2)\}$$

converges to a real, 2-D causal sequence. This causal sequence is the impulse response of the minimum-phase system whose log-magnitude (sampled) is $x(n_1, n_2)$.

Proof. From Lemma 4.2.9, Lemma 4.2.12 and Lemma 4.2.4, we know that

$$\mathcal{K}_c\{x(n_1, n_2)\} = \mathcal{K}_{C_1+C_2}\{x(n_1, n_2)\} \overset{2}{*} \mathcal{K}_{C_3}\{x(n_1, n_2)\}$$

where $\mathcal{K}_{C_1+C_2}\{x(n_1, n_2)\}$ and $\mathcal{K}_{C_3}\{x(n_1, n_2)\}$ are both real 2-D sequences. Therefore the convolution is also a real sequence. By Lemma 4.2.10, we know that $\mathcal{K}_C\{x(n_1, n_2)\}$ is a 2-D causal sequence. The fact that this sequence is the impulse response of a system whose log-magnitude (sampled) is $x(n_1, n_2)$ comes from the definition of $\mathcal{K}_C$. Since the first entry in this 2-D sequence is e^a, where a is the average value of $x(n_1, n_2)$, Theorem 3.4.4 tells us that this sequence is the impulse response of a minimum-phase system. ■

Hence we have shown that our method will produce the minimum-phase impulse response of a given log-magnitude in the limit as $N \to \infty$. In practice, since the convergence rate of this algorithm is $O(\frac{1}{N^2})$, a 512×512 transform size usually works very well. However, we have been able to find examples where a larger transform size is required to get the desired precision in our final answer.

4.3 History and Examples of the Extended Hilbert Transform Method

The first method that was attempted to find a minimum-phase system with a given log-magnitude response was the Planar Least Squares Inverse method, first introduced by Shanks [70, 35]. This method is an extension of a 1-D method for finding minimum-phase systems, and was conjectured to do the same for 2-D systems. A counter-example to this method was produced by Genin and Kamp [25, 26], and was subsequently proven to be false in the general case [26].

The second method that was attempted to find a minimum-phase system with a given log-magnitude was an extension of the Hilbert transform method that we have outlined in Section 1.6. The first example we have found of this method being attempted was by Read and Treitel [64], in 1973. Unfortunately, the causalizer used in this paper is incorrect, and the authors could not guarantee that their Hilbert transform method would produce a sequence that converged to a one-sided sequence. If we take $N = 4$, and reference Expression (4.1) again, a one-sided sequence by Read and Treitel's definition would be

$$\begin{array}{cccc}
a_{0,0\times} & a_{1,0} & 0 & 0 \\
a_{0,1} & a_{1,1} & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array}$$

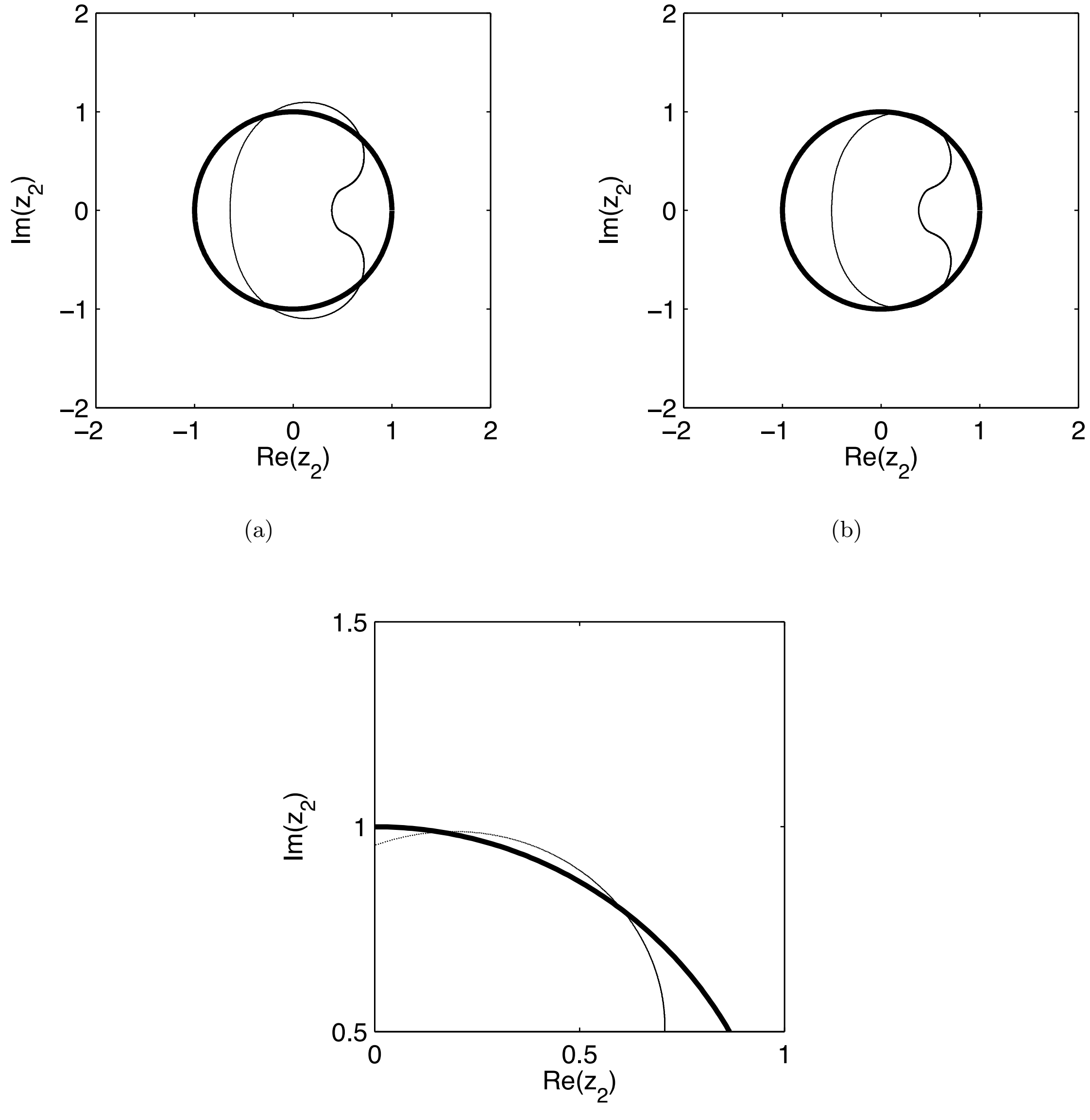
For the remainder of our discussion, this will be referred to as "quarter-plane causality", to contrast with our "half-plane causality".

In their paper, Read and Treitel introduced a log-magnitude whose impulse response, by their method, did not have a stable inverse. They considered the system with impulse response

$$\begin{array}{ccc}
1_\times & -1.20002759 & 0.40002239 \\
-1.00003018 & 1.70007079 & -0.65005088 \\
0.40002035 & -0.70054880 & 0.25004387
\end{array}$$

If we consider the transfer function of this system and plot the zeros for fixed values of z_1 where $|z_1| = 1$, we produce the root map given in Figure 4.1(a). The thick line represents the unit circle in the z_2-plane. Each point on the thin line represents a value of z_2 where $|z_1| = 1$ and the transfer function of this system evaluated at (z_1, z_2) is equal to zero. Note that there are values of z_2 that fall outside the unit circle. Therefore, by Huang's Theorem (Theorem 3.3.6), the system does not have a stable inverse and is not minimum-phase.

See Figure 4.1(b) for the rootmap of their resulting system. Notice that the roots slip slightly



(a)

(b)

(c)

Figure 4.1: (a) Rootmap of Read and Treitel's original system. (b) Rootmap of Read and Treitel's resulting system. (c) Resulting system zoomed in.

outside the unit circle. Originally, it was thought that this might be due to numerical errors in the computation, but this was later shown to be incorrect [10]. The main error in the derivation of Read and Treitel is due to their definition of quarter-plane causality in 2-D.

Our own analysis indicates that the method used by Read and Treitel produces a 2-D sequence which fails to converge to a causal sequence as the transform size grows. For a transform size of 512 × 512, we obtain the following output:

$$\begin{array}{ccccc}
-0.0001 & -0.0001 & 0.0000 & 0.0075 & -0.0032 \\
0.0000 & 0.0001 & 0.0002 & 0.0118 & -0.0068 \\
0.0002 & 0.0002 & 1.0000_{\times} & -1.1890 & 0.3868 \\
0.0048 & -0.0001 & -0.9998 & 1.6584 & -0.6269 \\
-0.0032 & -0.0015 & 0.3837 & -0.6731 & 0.2414
\end{array}$$

(This output has been normalized so that the entry at the origin is exactly 1.) The output is actually a much larger sequence, but showing the values around the origin shows enough information to make our point; note the entries outside of the quarter-plane region specified by Read and Treitel. In fact, they are outside the half-plane region required for half-plane causality. According to Read and Treitel, any entry to the left , or above, the origin would be a non-causal entry. Now, if we try a transform size of 1024 × 1024, we obtain:

$$\begin{array}{ccccc}
-0.0001 & -0.0001 & 0.0000 & 0.0075 & -0.0032 \\
0.0000 & 0.0001 & 0.0002 & 0.0118 & -0.0068 \\
0.0002 & 0.0002 & 1.0000_{\times} & -1.1890 & 0.3868 \\
0.0048 & -0.0001 & -0.9998 & 1.6584 & -0.6269 \\
-0.0032 & -0.0015 & 0.3837 & -0.6731 & 0.2414
\end{array}$$

The output is unchanged, and still non-causal (either quarter-plane or half-plane causal). This suggests that the output is not converging, and that the non-causal entries are significant.

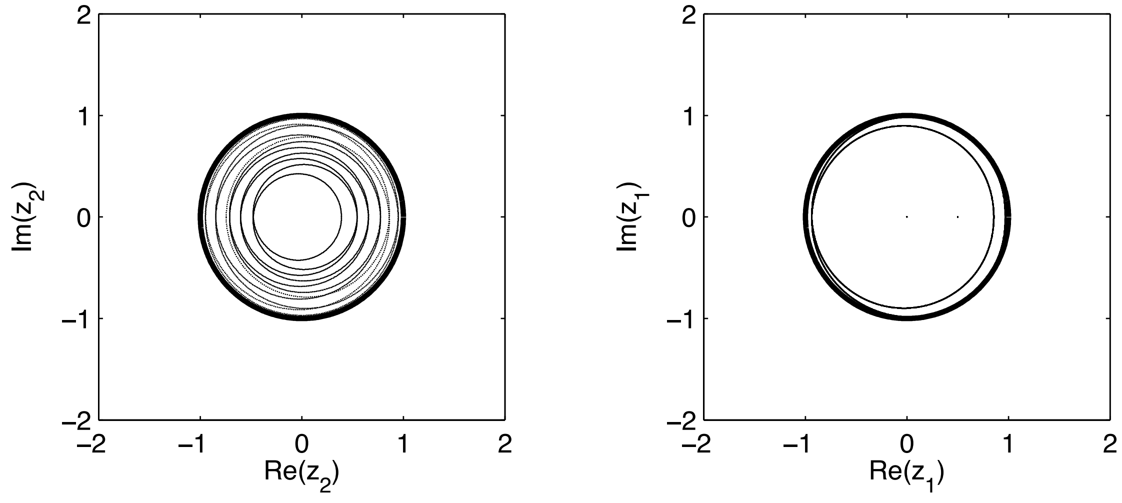
According to Bose [10], the method used by Read and Treitel was incorrect due to an inconsistency with the factorization problem. That is, for a given amplitude function squared, $A^2(i\omega_1, i\omega_2) = |H(i\omega_1, i\omega_2)|^2$, where H is the transfer function of a minimum-phase system (in the sense of Read and Treitel), we cannot necessarily factor A^2 into $A^2(i\omega_1, i\omega_2) = H(p_1, p_2)H(-p_1, p_2)$, where $\omega_1^2 = -p_1^2$ and $\omega_2^2 = -p_2^2$. (See the article by Bose [10] for details.) The method presented in this chapter circumvents this problem by eliminating the restriction that the transfer function of a minimum-phase system only contains negative powers of z_1 and z_2.

By taking the log-magnitude of the original system, we can use our method to find the correct minimum-phase system. Using a 512 × 512 transform, we obtain the impulse response

0.0000	0.0000	−0.0000	−0.0000	−0.0000	0.0000	0.0000
0.0000	−0.0000	−0.0000	−0.0000	0.0000	0.0000	0.0000
−0.0000	−0.0000	−0.0000	$1.0000_\times$	−1.1743	0.3743	−0.0116
0.0125	0.0097	0.0001	−1.0108	1.6583	−0.6234	0.0062
−0.0058	−0.0065	−0.0030	0.4028	−0.6825	0.2416	0.0000

This impulse response has been normalized so that the entry at the origin is 1. Unlike Read and Treitel's scheme, our output is half-plane causal (to at least 4 decimal places). The output of our method is identical if we use a transform size of 1024 × 1024. So we can safely assume that this impulse response is accurate to 4 decimal places. Notice that the values of this impulse response and that of the nonminimum-phase system are very close, but the difference between them is significant. There are non-zero causal entries outside the range of values shown here. That is, what is shown is not the complete minimum-phase impulse response. It is a truncated version of the minimum-phase impulse response.

See Figure 4.2 (a) to see the rootmap of the related quarter-plane system while holding z_1 fixed. Also, see Figure (4.2 (b)) to see the rootmap of the related quarter-plane system while holding z_2 fixed. Note that all of the roots lie inside the unit circle. Therefore, this system is minimum-phase by Huang's theorem.



(a)

(b)

Figure 4.2: (a) Rootmap of Read and Treitel's counter-example using our method, holding z_1 fixed. (b) Rootmap of Read and Treitel's counter-example using our method, holding z_2 fixed.

At this point, there was an investigation to find under what circumstances the Read and

Treitel Hilbert transform method would produce a minimum-phase system. It has been shown [58] that, for a given transform size, the 1-D Hilbert transform method does not necessarily give the impulse response of a minimum-phase system. Problems may occur if the transfer function of the system has zeros on the unit circle. The Hilbert transform will still converge to the correct minimum-phase system as the transform size grows. Murthy and Reddy [57] conjectured that this was the reason for the failure of Read and Treitel's method in the counterexample above, since this system does have zeros on the unit bidisc, $|z_1| = |z_2| = 1$. They also suggested that if the original system has a transfer function without zeros on the unit polydisc (equivalently, the log-magnitude of the system is bounded), then Read and Treitel's method will produce the transfer function of a minimum-phase system with the same log-magnitude.

For our next example, we will use the transfer function from Example 3.3.8, $1 - \frac{1}{2} z_1 z_2^{-1}$. This transfer function does not have any zeros on the unit polydisc, and therefore, as conjectured by Murthy and Reddy, Read and Treitel's method should converge to a minimum-phase impulse response.

The results from this test are surprising. In Example 3.3.8, we were able to show that this system is minimum-phase by explicitly computing the inverse system and summing its impulse response. However, this is not a quarter-plane system, and the definition of causality used by Read and Treitel suggests that their method should produce a minimum-phase quarter-plane system. The actual output of the Read and Treitel method converges to:

$$\begin{array}{ccccc}
0 & 0 & 0 & 0 & -0.0275 \\
0 & 0 & 0 & -0.2275 & 0 \\
0 & 0 & 1_\times & 0 & 0 \\
0 & -0.2275 & 0 & 0 & 0 \\
-0.0275 & 0 & 0 & 0 & 0
\end{array}$$

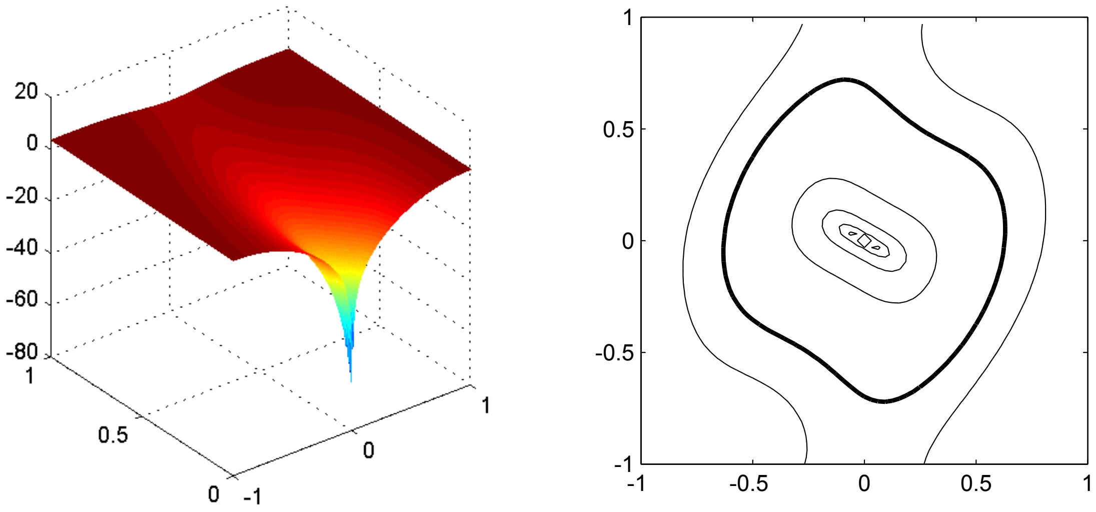
(This only shows a few values of the impulse response around the origin). This impulse response is even, and definitely not causal, by either the quarter-plane or half-plane definition. In fact, the 1 at the origin is the only non-zero value in the causal region as defined by Read and Treitel.

For comparison, our method returns the following impulse response:

$$\begin{array}{ccc}
0 & 0 & 0 \\
0 & 1_\times & 0 \\
-0.5 & 0 & 0
\end{array}$$

which we know to be the correct solution.

Now that we have gone through some of the history of designing minimum-phase systems, what we would like to do with our extended Hilbert transform method is to create a noise shaper that improves on the benchmark Floyd and Steinberg shaper. Recall Section 2.4. The Floyd and Steinberg shaper is minimum-phase, and its log-magnitude can be seen in Figure 4.3. The vertical scale is in decibels, and the surfaces are plotted as functions of normalized frequency (1 represents the Nyquist frequency). From Figure 2.18, we can see that the Floyd and Steinberg shaper does a reasonably good job at producing a quality image. Figure 4.3 explains why. The low frequency noise is very heavily suppressed. Note that only half of the surface is shown in Figure 4.3(a) since the surface is even. All of the information not included in this surface can be extrapolated, and the sharp peak at the origin (the dc point) is more visible than if the entire surface was shown. Note that the surface has equal volumes above and below the 0 dB plane. This is due to the fact that the Floyd and Steinberg design is minimum-phase.



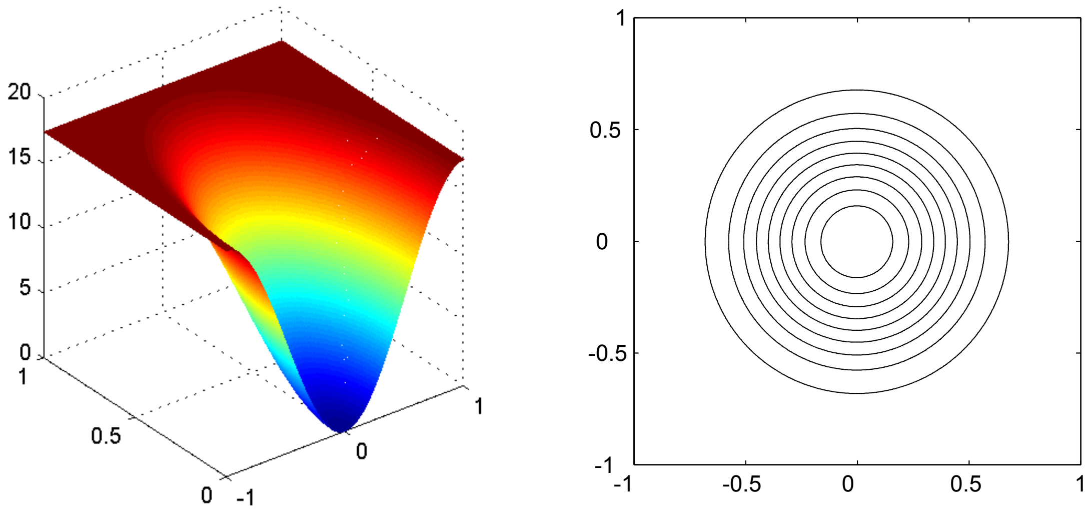
(a)

(b)

Figure 4.3: (a) Log-magnitude of the Floyd and Steinberg shaper. Only half of the frequency response is shown since the surface is even. Plotted in spatial frequency vs. decibels. (b)Contour plot of the log-magnitude. The heavy line denotes the contour of amplitude 0dB.

Recall Figure 2.2, the visibility curve. If we assume radial symmetry of the sensitivity of the human eye to spatial frequency, then we can see that the log-magnitude of the Floyd and Steinberg system is a poor fit to the visibility surface. There are a couple of things we might do to improve upon the design of the Floyd and Steinberg shaper. We will design a shaper that is radially symmetric and does not have any sharp nulls. Recall the Extended Jensen's Inequality: the sharp null in the Floyd and Steinberg shaper is a large volume that we could distribute elsewhere to suppress a larger range of frequencies, instead of a large suppression of a small range of frequencies. Our target shape will be a raised negative cosine. This surface gives a reasonable agreement to the inverted visibility curve, Figure 2.2, assuming radial symmetry

(See Figure 4.4). Notice that the log-magnitude does not have equal volumes above and below the zero plane. The shaper we end up using will have the same shape, but will be translated down so that we have equal volumes above and below 0 dB.



(a)

(b)

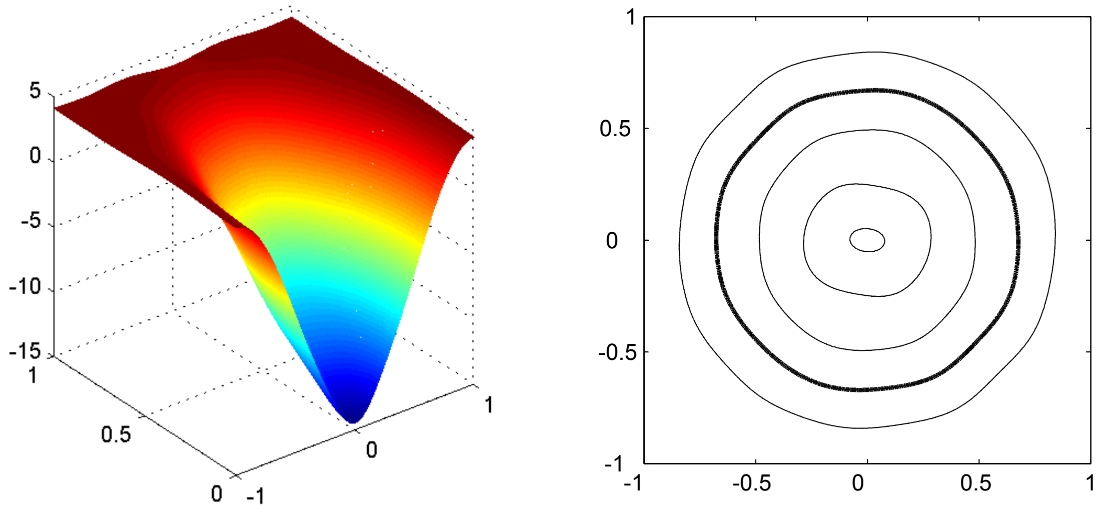
Figure 4.4: (a) The target log-magnitude shape, plotted in spatial frequency vs. decibels. Only half of the frequency response is shown since the surface is even. (b) Contour plot of the log-magnitude surface.

Using a 512 × 512 transform, we obtain the following output:

$$\begin{array}{ccccccc}
 & & & 1_{\times} & -0.5090 & 0.1008 & -0.0009 \\
0.0015 & 0.0057 & -0.2549 & -0.3802 & -0.0180 & 0.0834 & -0.0255 \\
-0.0082 & 0.0447 & 0.1114 & 0.1007 & 0.0627 & -0.0106 & -0.0154 \\
-0.0035 & -0.0256 & -0.0244 & -0.0193 & -0.0234 & -0.0111 & 0.0077
\end{array}$$

These are just the first few entries around the origin. All of the non-causal entries are zero to 4 decimal places, and will converge to zero if we increase the transform size. The remaining entries are correct to at least 3 decimal places. The entries outside the range shown have been truncated, which means there will be some error between the log-magnitude of this shaper, and the target log-magnitude shown in Figure 4.4. The log-magnitude of this shaper can be seen in Figure 4.5. We can see that there is very close agreement in the shape of these two log-magnitudes. Note that the log-magnitude in Figure 4.5 has equal volumes above and below the 0 dB plane, as expected for a minimum-phase system of the form $1 - H(z_1, z_2)$.

Now, to see how well this shaper performs, see Figure 4.6. The error of this signal can be found in Figure 4.7. Recall that these images are quantized to only 3 bits per colour, a very



(a)

(b)

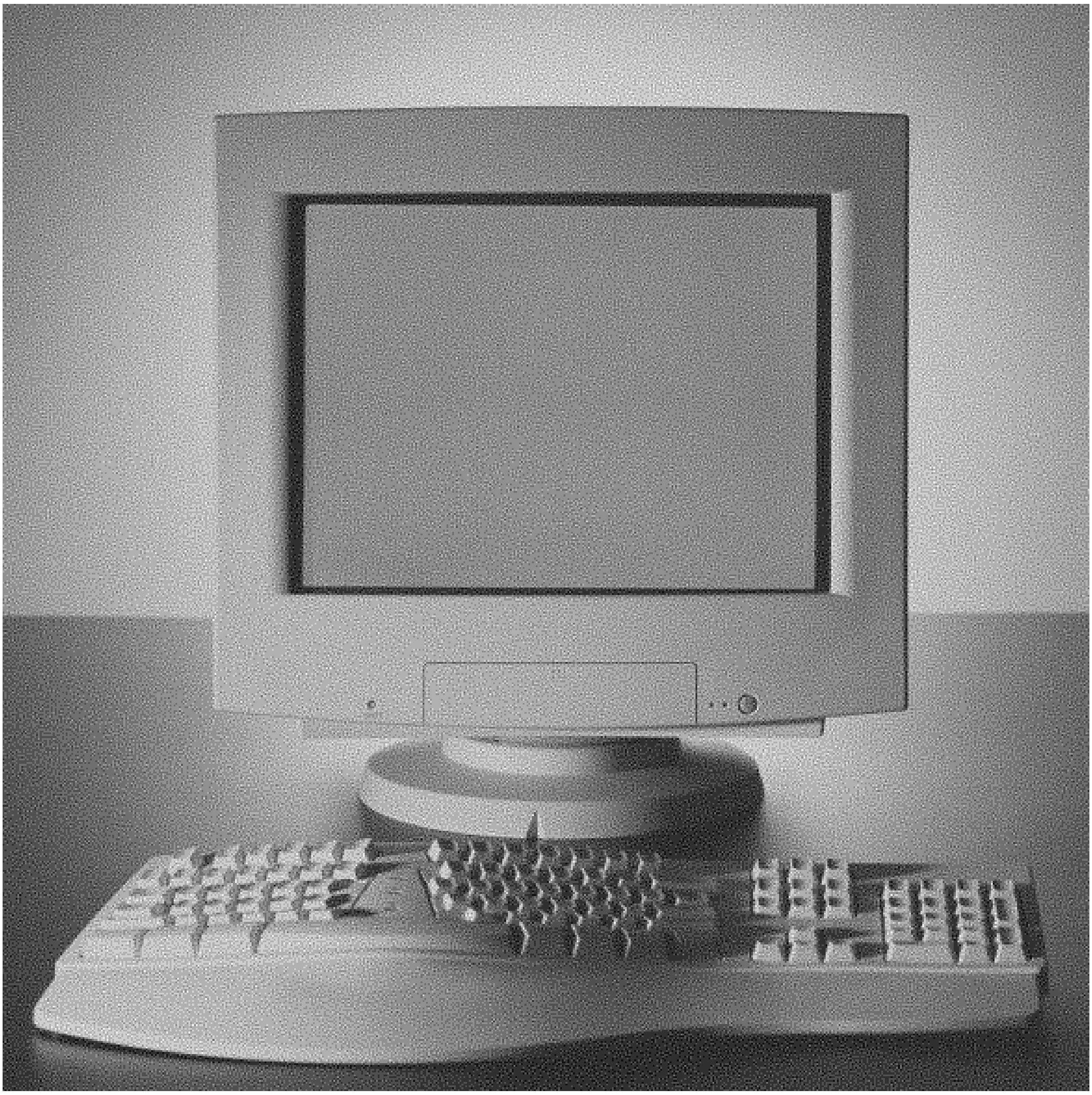
Figure 4.5: (a) The log-magnitude of the designed shaper, plotted in normalized spatial frequency vs. decibels. (b) Contour plot of the log-magnitude. The heavy line indicates the contour of amplitude 0dB.

coarse quantization for images. These results, in the author's opinion, offer a slight improvement over the image produced by the Floyd and Steinberg shaper, Figure 2.18. The image appears to have less low frequency noise than the image using the Floyd and Steinberg shaper. This is due to the fact that we designed this shaper to suppress a larger range of frequencies around dc.

4.4 Conclusions

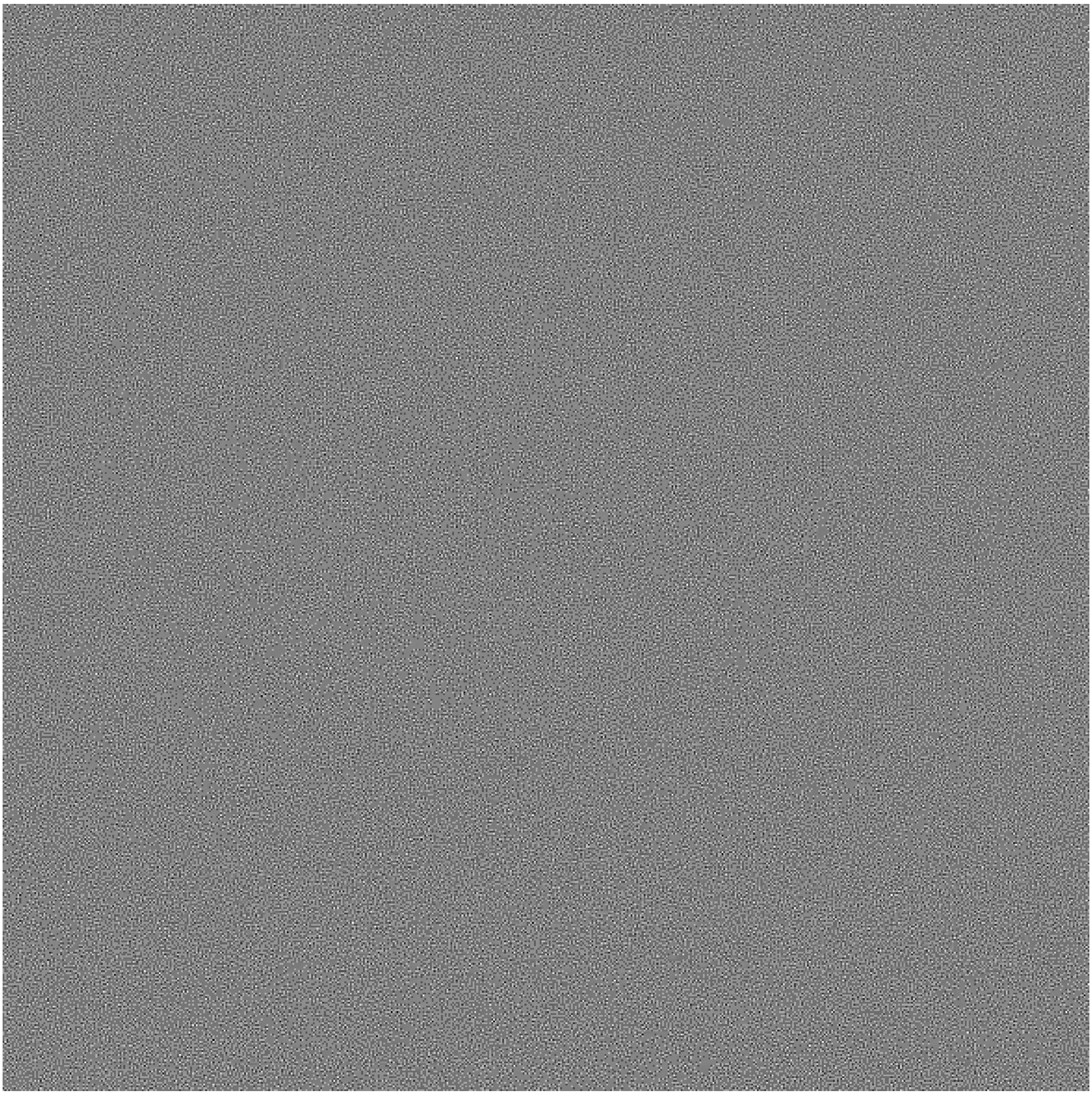
The method that we have outlined here can be used to find a minimum-phase system with almost any log-magnitude shape desired. In all tests that were conducted, the output impulse response converged at a rate of at least $O(\frac{1}{N^2})$. This includes all counter-examples we could find in the literature that showed how other versions of the extended Hilbert transform method can fail. Additionally, we have offered a proof of why our extended Hilbert Transform method works; we have not been able to find a proof offered for any other version of the extended Hilbert Transform method in the literature.

It was mentioned that the minimum-phase noise shaper developed in this chapter offers only a slight improvement over the Floyd and Steinberg shaper. We can create noise shapers with more drastic improvements over the Floyd and Steinberg design, but that would require more headroom in our target image to ensure we do not saturate the quantizer; greater noise shaping comes with a greater headroom requirement. In practice, better noise shaping results could be obtained if we used a less coarse quantization of the original image, or if we interpolate the



(a)

Figure 4.6: Noise shaped image using TPDF dither and negative cosine shaped log-magnitude.



(a)

Figure 4.7: The error of the noise shaped image.

sampled data points (oversampling).

The methods presented in this section are readily extensible to higher dimensions. That is to say, we can use this method to design minimum-phase systems of any dimension, and not just the 1- and 2-dimensional systems discussed in this thesis.

Chapter 5

Conclusions and Considerations

In this thesis, we have been able to demonstrate how dithered quantization and noise shaping apply to digital images as an extension of the same concepts from digital audio. We have shown the results of applying different dithered quantization techniques to digital images, and discussed the positive and negative contributions to the quantization problem. However, there are still many questions that can be explored from this work.

The question of dithered quantization and noise shaping of colour images requires further investigation. This theory would be more analogous to quantization of stereo signals in digital audio. The methods presented in this thesis are not optimal when applied to colour images, but we have seen that when applied in a straightforward fashion, the results are reasonably good. The author is confident that further investigation in this area can yield even better results.

The two most important mathematical contributions of this thesis are the extensions of Jensen's Inequality and the discrete Hilbert transform to higher (finite) dimensions. Aside from its application to digital imaging, the extended Jensen's Inequality may yield new results in the theory of functions of several complex variables, and is of interest to anyone in the field of multi-dimensional signal processing.

The extension of the discrete Hilbert transform to higher dimensional signals will also be of use in multi-dimensional signal processing. This method has direct applications to geophysicists, who were likely the first to study the problem of minimum-phase systems in a higher dimensional setting. From a mathematical perspective, it would be useful to reverse engineer a 2-D form of the Hilbert transform, and then see if alternative derivations exist similar to the development of the 1-D Hilbert transform; that is, using methods of contour integration, and symmetry properties of odd and even functions. A deeper investigation of a 2-D Hilbert transform, coupled with the extended version of Jensen's Inequality, may lead to new understandings of the analyticity properties of functions of several complex variables, and the locations of their zero surfaces.

Bibliography

[1] L.V. Ahlfors. *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable.* McGraw-Hill Book Company, second edition, 1966.

[2] J.P. Allebach and B. Liu. Analysis of halftone dot profile and aliasing in the discreate binary representation of images. *Journal of the Optical Society of America*, 67(9):1147–1154, September 1977.

[3] D. Anastassiou. Error diffusion coding for a/d conversion. *IEEE Transactions on Circuits and Systems*, 36(9):1175–1186, September 1989.

[4] D. Anastassiou and K.S. Pennington. Digital halftoning of images. *IBM Journal of Research and Development*, 26(6):687–697, November 1982.

[5] A.S. Willsky A.V. Oppenheim and S.H. Nawab. *Signals and Systems.* Prentice Hall, second edition, 1997.

[6] R.W. Schafer A.V. Oppenheim and J.R. Buck. *Discrete-Time Signal Processing.* Prentice Hall, second edition, 1999.

[7] N. Bekkat and A. Saadane. Coded image quality assessment based on a new contrast masking model. *Journal of Electronic Imaging*, 13(2):341–348, April 2004.

[8] A.J. Berkhout. On the minimum phase criterion of sampled signals. *IEEE Transactions on Geoscience Electronics*, 11:186–198, 1973.

[9] C. Billotet-Hoffmann and O. Bryngdahl. On the error diffusion technique for electronic halftoning. *Proceedings of the S.I.D.*, 24(3):253–258, 1983.

[10] N.K. Bose. Problems in stabilization of multidimensional filters via Hilbert transform. *IEEE Transactions of Geoscience Electronics*, pages 146–147, 1974.

[11] L.K. Brinton. Nonsubtractive dither. Master's thesis, University of Utah, 1984.

[12] Z.L. Budrikis. Visual fidelity criterion and modeling. *Proceedings of the IEEE*, 60(7):771–779, July 1972.

[13] S. Chakrabarti and S.K. Mitra. Design of two-dimensional digital filters via spectral transformations. *Proceedings of the IEEE*, 65(6):905–914, June 1977.

[14] C.H. Chen and C.Y. Chi. Statistical texture image classification using two-dimensional nonminimum-phase fourier series based model. *Proceedings of the IEEE Signal Processing Workshop*, pages 400–403, 1999.

[15] R. V. Churchill. *Complex variables and applications.* New York: McGraw-Hill, and Tokyo: Kogakusha Comp., 1960, 2nd ed., 1960.

[16] J.F. Jarvis C.N. Judice and W.H. Ninke. Using ordered dither to display continuous tone pictures on an ac plasma panel. *Proceeding of the S.I.D.*, 15(4):161–169, 1974.

[17] S. Colombo. *Holomorphic Functions of One Variable.* Gordon and Breach Science Publishers, first edition, 1983.

[18] D.E. Dudgeon. Fundamentals of digital array processing. *Proceedings of the IEEE*, 65(6):898–904, June 1977.

[19] V.R. Kolavennu E.I. Jury and B.D.O. Anderson. Stabilization of certain two-dimensional recursive digital filters. *Proceedings of the IEEE*, 65(6):887–892, June 1977.

[20] M.P. Ekstrom and J.W. Woods. Two-dimensional spectral factorization with applications in recursive digital filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(2):115–128, April 1976.

[21] R. Eschbach. Error diffusion algorithm with homogeneous response in highlight and shadow areas. *Journal of Electronic Imaging*, 6(3):348–356, July 1997.

[22] S. Weissbach F. Fetthauer and O. Bryngdahl. Optimization of error diffusion filter for design of digital phase hologram. *Optics Communications*, 94:44–48, November 1992.

[23] S. Weissbach F. Fetthauer and O. Bryngdahl. Equivalence of error diffusion and minimal average error algorithms. *Optics Communications*, 113:365–370, January 1995.

[24] R.W. Floyd and L. Steinberg. An adaptive algorithm for spatial greyscale. *Proceeding of the S.I.D.*, 17(2):75–77, 1976.

[25] Y.V. Genin and Y.G. Kamp. Counterexample in the least-squares inverse stabilisation of 2d recursive filters. *IEEE Electronic Letters*, 11(15):330–331, July 1975.

[26] Y.V. Genin and Y.G. Kamp. Two-dimensional stability and orthogonal polynomials on the hypercircle. *Proceedings of the IEEE*, 65(6):873–881, June 1977.

[27] S.L. Hahn. *Hilbert Transforms in Signal Processing.* Artech House, Norwood, Massachusetts, first edition, 1996.

[28] W.Y. Han and J.C. Lin. Error diffusion without contouring effect. *Journal of Electronic Imaging*, 6(1):133–139, January 1997.

[29] A.C. Bovik H.R. Sheikh and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, December 2005.

[30] T.S. Huang, editor. *Picture Processing and Digital Filtering*, volume 6 of *Topics in Applied Physics*. Springer-Verlag, Berlin, second edition, 1979.

[31] R.J. Marks II, editor. *Introduction to Shannon Sampling and Interpolation Theory.* Springer-Verlag, New York, 1991.

[32] A. Jeffrey. *Complex Analysis and Applications.* CRC Press, Boca Raton, Florida, first edition, 1992.

[33] J.L.V.W. Jensen. Sur un nouvel et important thèorème de la thèorie des fonctions. *Acta Mathematica*, 22(1):359–364, December 1899.

[34] C.N. Judice J.F. Jarvis and W.H. Ninke. A survey of techniques for the display of continuous tone pictures on bilevel displays. *Computer Graphics and Image Processing*, 5:13–40, 1976.

[35] S. Treitel J.L. Shanks and J.H. Justice. Stability and synthesis of two-dimensional recursive filters. *IEEE Transactions on Audio and Electroacoustics*, 20(2):115–128, June 1972.

[36] J.L.Shanks. Recursion filters for digital processing. *Geophysics*, 32(1):33–51, February 1967.

[37] E.I. Jury. Stability of multidimensional scalar and matrix polynomials. *Proceedings of the IEEE*, 66(9):1018–1047, September 1978.

[38] J.H. Justice and J.L. Shanks. Stability criterion for n-dimensional digital filters. *IEEE Transactions on Automatic Control*, 18:284–286, June 1973.

[39] D. Kermisch and P.G. Roetling. Fourier spectrum of halftone images. *Journal of the Optical Society of America*, 65(6):716–723, June 1975.

[40] K.T. Knox. Error image in error diffusion. *Image Processing Algorithms and Techniques*, pages 268–279, 1992.

[41] K.T. Knox. Error diffusion: A theoretical view. *SPIE*, 1913:326–331, 1993.

[42] K.T. Knox. Evolution of error diffusion. *Journal of Electronic Imaging*, 8(4):422–429, October 1999.

[43] K.T. Knox and R. Eschbach. Threshold modulation in error diffusion. *Journal of Electronic Imaging*, 2(3):185–192, July 1993.

[44] D.E. Knuth. Digital halftones by dot diffusion. *ACM Transactions on Graphics*, 6(4):245–273, October 1987.

[45] B.W. Kolpatzik and C.A. Bouman. Optimized error diffusion for image display. *Journal of Electronic Imaging*, 1(3):277–292, July 1992.

[46] J.S. Lim and A.V. Oppenheim, editors. *Advanced Topics in Signal Processing.* Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey, first edition, 1988.

[47] B. Lippel and M. Kurland. The effect of dither on luminance quantization of pictures. *IEEE Transactions on Communication Technology*, 19(6):879–888, December 1971.

[48] S.P. Lipshitz and C.N. Christou. Picturing dither: Dithering pictures. Presented at the 121st Convention of the Audio Engineering Society, San Francisco, October 2006.

[49] J.L. Mannos and D.J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, July 1974.

[50] M. Marden. *The Geometry of the Zeros of a Polynomial in a Complex Variable.* American Mathematical Society, New York, first edition, 1949.

[51] J.E. Marsden. *Basic Complex Analysis.* W.H. Freeman and Company, 1973.

[52] T.L. Marzetta. Two-dimensional linear prediction: Autocorrelation arrays, minimum-phase prediction error filters, and reflection coefficient arrays. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(6):725–733, December 1980.

[53] T.L. Marzetta. Additive and multiplicative minimum-phase decompositions of 2-d rational power density spectra. *IEEE Transactions on Circuits and Systems*, 29(4):207–214, April 1982.

[54] T.L. Marzetta. The minimum energy-delay property of 2-d minimum-phase filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(4):658–659, August 1982.

[55] S.K. Mitra and M.P. Ekstrom, editors. *Two-Dimensional Digital Signal Processing*, volume 20 of *Benchmark Papers in Electrical Engineering and Computer Science.* Dowden, Hutchinson & Ross, Inc., first edition, 1978.

[56] T. Mitsa and K.J. Parker. Digital halftoning technique using a blue-noise mask. *Journal of the Optical Society of America*, 9(11):1920–1929, November 1992.

[57] H.S.N. Murthy and P.S. Reddy. On the proof of stabilization of two-dimensional recursive filters via discrete Hilbert transform. *IEEE Transactions on Circuits and Systems*, 33(8):741–749, August 1986.

[58] M.S. Hrishikesh P.S. Reddy N. Damera-Venkata, M. Venkataraman. Stabilization of 2-d recursive digital filters by the dht method. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, 46(1):85–88, January 1999.

[59] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing.* Prentice-Hall, New Jersey, first edition, 1975.

[60] A. Papoulis, editor. *The Fourier Integral and its Applications.* McGraw-Hill Electronic Sciences Series. mcGraw-Hill, first edition, 1962.

[61] A.D. Poularikas, editor. *The Transforms and Applications Handbook.* The Electrical Engineering Handbook Series. CRC Press, second edition, 2000.

[62] J. Vanderkooy J.N. Wright R.A. Wannamaker, S.P. Lipshitz. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing*, 48(2):499–516, February 2000.

[63] L.R. Rabiner and B. Gold. *Theory and Application of Digital Signal Processing.* Prentice Hall, first edition, 1975.

[64] R.R. Read and S. Treitel. The stabilization of two-dimensional recursive filters via the discrete Hilbert transform. *IEEE Trans. Geoscience Electronics*, pages 153–160, 1973.

[65] R.J. Rolleston and S.J. Cohen. Halftoning with random correlated noise. *Journal of Electronic Imaging*, 1(2):209–217, April 1992.

[66] J.L. Shanks S. Treitel and C.W. Frasier. Some aspects of fan filtering. *Geophysics*, 32(5):789–800, October 1967.

[67] M.R. Schroeder. Images from computers. *IEEE Spectrum*, pages 66–78, March 1969.

[68] M.R. Schroeder. Images from computers and microfilm plotters. *Communications of the ACM*, 12(2):95–101, February 1969.

[69] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, pages 162–165, December 1964.

[70] J.L. Shanks. Two planar digital filtering algorithms. *Princeton Conference on Information Sciences and Systems*, 5:48–53, March 1971.

[71] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, February 2006.

[72] D.T. Sherwood. Some theorems on quantization and an example using dither. *Conference Record. Nineteenth Asilomar Conference on Circuits, Systems and Computers*, pages 207–212, 1986.

[73] R.A. Wannamaker S.P. Lipshitz and J. Vanderkooy. Quantization and dither: A theoretical survey. *Journal of the Audio Engineering Society*, 40(5):355–375, May 1992.

[74] H.A. Spang and P.M. Schultheiss. Reduction of quantizing noise by use of feedback. *IRE Transactions on Communications Systems*, 10(4):373–380, December 1962.

[75] 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*, 15(5):442–448, October 1977.

[76] S. Weissbach T. Zeggel and O. Bryngdahl. Noise modification in iterative multi-level image quantization. *Optics Communications*, 100:67–71, 1993.

[77] S. Treitel and J.L. Shanks. The design of multistage separable planar filters. *IEEE Transactions on Geoscience Electronics*, 9(1):10–27, January 1971.

[78] V. Čížek. Discrete Hilbert transform. *IEEE Transactions on Audio and Electroacoustics*, 81(4):340–343, December 1970.

[79] R. Ulichney. *Digital Halftoning.* MIT Press, Cambridge, Massachusetts, first edition, 1987.

[80] R.A. Ulichney. Dithering with blue noise. *Proceedings of the IEEE*, 76(1):56–79, January 1988.

[81] R.A. Wannamaker. Dither and noise shaping in audio applications. Master's thesis, University of Waterloo, Waterloo, Ontario, 1991.

[82] R.A. Wannamaker. Psychoacoustically optimal noise shaping. *Journal of the Audio Engineering Society*, 40(7):611–620, July/August 1992.

[83] R.A. Wannamaker. *The Theory of Dithered Quantization.* PhD thesis, University of Waterloo, Waterloo, Ontario, 1997.

[84] S. Weissbach and O. Brygndahl. Control of halftone texture by error diffusion. *Optics Communications*, 103:174–180, 1993.

[85] S. Weissbach and F. Wyrowski. Error diffusion procedure: Theory and applications in optical signal processing. *Applied Optics*, 31(14):2518–2534, May 1992.

[86] S. Weissbach and F. Wyrowski. Numerical stability of the error diffusion concept. *Optics Communications*, pages 151–155, 1992.

[87] E.T. Whittaker. On the functions which are represented by the expansions of the interpolation-theory. *Proceedings of the Royal Society of Edinburgh*, 35:181–194, 1915.

[88] S.K.M. Wang W.K. Chau and S.J. Wan. A critical analysis of dithering algorithms for image processing. *IEEE Region 10 Conference on Computer and Communications Systems*, pages 309–313, September 1991.