ADAPTIVE DITHERING OF ONE DIMENSIONAL SIGNALS

By

Carlos Fabian Benitez-Quiroz

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

2007

Approved by:

Damaris Santana, Ph.D Member, Graduate Committee	Date
Miguel Velez-Reyes, Ph.D Member, Graduate Committee	Date
Shawn D. Hunt, Ph.D President, Graduate Committee	Date
Luis F. Cáceres, Ph.D Representative of Graduate Studies	Date
Isidoro Couvertier, Ph.D Chairperson of the Department	Date

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

ADAPTIVE DITHERING OF ONE DIMENSIONAL SIGNALS

By

Carlos Fabian Benitez-Quiroz

2007

Chair: Shawn D. Hunt
Major Department: Electrical and Computer Engineering

This research work presents new techniques on the dithering of one dimensional signals. The first part of this work was to determine whether dither is needed when reducing the bit depth. This was done with time series hypothesis tests in time and frequency domains, and they evaluate whether the total error of the undithered quantization is white noise. If this is the case, then no undesired harmonics are added in the quantization or re-quantization process. The second part of this work was to develop a state of the art dither which is signal dependent and attempts to have lower quantization noise than the classic techniques. Experiments showing the effectiveness of the methods with both synthetic and real audio signals are presented.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

DITHERING ADAPTIVO EN SEÑALES DE UNA DIMENSION

Por

Carlos Fabian Benitez-Quiroz

2007

Consejero: Shawn D. Hunt
Departamento: Departamento

En este trabajo de investigacion se muestran nuevas técnicas sobre dithering en señales de una dimensión. La primera parte es determinar si dithering es necesario cuando se reduce el número de bits en una señal digital. Esto se determina usando pruebas de hipotesis en series de tiempo en el dominio del tiempo y de la frecuencia, en las cuales se evalua si el error total de cuantización es ruido blanco. La segunda parte de esta investigación fue desarrollar un dither que se encuentra en el estado del arte el cual sea dependiente de la señal y tengo un ruido de cuantización menor que el de las técnicas clásicas. Experimentos realizados en esta investigación muestra la eficiencia de los métodos en audio sintético y real.

To my parents Ciro Benitez and Alma Quiroz, my brothers Javier and Ciro, and my nephews Javier Alejandro and Juan Sebastian .

To Carolina, for her support and love during my research at the University of Puerto Rico and for encourage me to finish this research.

ACKNOWLEDGMENTS

I want to start expressing my most truthful gratitude and acknowledgement to my advisor, Dr. Shawn Hunt for encouraging me to work in this research and providing assistance and support for this work. From him, I received excellent education and many talks about different researches and many advices about life. Without his guidance this work could not have been finished.

I would also like to thank Professor Dr. Damaris Santana for helping me in statistical theory and giving me information about books, journals and advising me in my career in the future. I want to give acknowledgments to Professors Pedro Vazquez, Edgar Acuña and Domingo Rodriguez for helping me in optimization problems, hypothesis testing and circular deconvolution problems respectively. I would also like to thank Dr. Miguel Velez for having been in my committee.

I specially would like to thank Luis Quintero (a.k.a Checho) for encouraging me to work in the days when I did not find solutions, and discussing with me topics about the research and LaTeX. Also, I would like to thank Professor Miguel Figueroa for helping me find optimization package and advising me in grammar and in the presentation of the results of this research. I want to express gratitude to Jessica Jimenez for helping me in the writing of this document and to Maria Isabel Santacoloma for reviewing my thesis in many opportunities. To Angela Anaya for helping me writing and review this work. I want acknowledgments to go to the graduate students of the INEL/ICOM department and to my friends who are near my work, such as Carlos and Diego Aponte, Lola Bautista, Felix Mar Luna, Carlos Niño Baron, Fernando Cintron, Laura Sanchez, Diego Arias, Natalia Gonzales and Ana Ramirez. To my house mates in Puerto Rico (almost my family) Maria

Fernanda Naranjo, Hector Figueroa, Carolina Peña, Carolina Gerardino, Shelly, Gerardo and Maria Helena Torres. I would like to thank the administrative personal of INEL and CenSSIS such as Sandy, Keyla, Maribel, Claribel, Vanessa and Markus.

Finally I would like to thank God for giving me the opportunity to study at this prestigious University.

This work was supported by CenSSIS, Bernard Gordon Center for Subsurface Sensing and Imaging Systems, under the Engineering Research Centers Program of the National Science Foundation (Award Number EEC-998621).

TABLE OF CONTENTS

page

LIST OF TABLES

Table page

LIST OF FIGURES

LIST OF ABBREVIATIONS

FFT	Fast Fourier Transform.
DCFT	Discrete Chirp Fourier Transform.
AR	Autoregressive.
MA	Moving Average.
ARMA	Autoregressive Moving Average.
PDF	Probability density function.
RPDF	Uniform density function.
GPDF	Gaussian density function.
SNR	Signal to noise ratio.
i.i.d	Independent and identically distributed.
PCM	Pulse Code Modulation.
NSD	Non subtractive dither.
SD	Subtractive dither.
UD	Undithered.
ANSD	Adaptive Nonsubtractive dither.
SDD	Segment dependent dither.
ND	Numerical dither.
SPG	Spectral Projected Gradient.
LM	Levenberg-Marquardt.

LIST OF SYMBOLS

t	Time (seconds).
Hz	Frequency (Hertz).

CHAPTER 1
INTRODUCTION

Digital signals have become widespread and are favored in many applications because of their noise immunity. This is a result of their discreteness in both time and amplitude. Once the signal has been discretized, the signal can be stored or transmitted without additional noise being added. There are many applications however, where the discretization in both time and amplitude needs to be changed in the discrete domain. For example, if two digital systems operate at different sampling frequencies, a multirate system is used for the sampling rate conversion. The amplitude discretization can also be changed. Going from a 16 to an 8 bit representation, for example, would reduce the memory requirements for storing a signal. For instance, in gaming applications the precision of each sample is about 12 bits and the ring tones have a resolution of 8bits. The process of lowering the amplitude resolution of a digital signal is called re-quantization.

If the signal amplitude change is large from sample to sample, then it is generally assumed that the re-quantization will be a uniformly (discrete) distributed i.i.d. (independent and identically distributed) sequence (white noise). In this case, the re-quantization error is independent of the signal being quantized. This assumption does not hold for all cases particularly when the signal has small amplitude. In this case rounding or truncating a signal can introduce various undesirable artifacts, namely, additional harmonics related to the signal being re-quantized. For this case, the re-quantization error is an autoregressive moving average ($ARMA$) process that can not be modeled like a white noise sequence.

To avoid these unwanted harmonics, dither is generally added to the signal being quantized. Dither is an i.i.d. signal whose purpose is to ensure that the quantization error is uncorrelated with the signal being quantized. In addition, the dither signal is independent of the input. In additive dithering, the quantization error signal is dependent of the signal being quantized but some techniques can make the value of the first and second statistical moments independent of the error. The main disadvantage of adding dither is that since the dither is basically a noise signal, the signal to noise ratio (SNR) of the final re-quantized signal is lowered. Because of this, it is desirable to know if re-quantization will introduce the undesired harmonics. In some cases, the quantization noise will be an i.i.d signal even though no dither is added. In these cases, the signal can be re-quantized with no added harmonics, and without the signal to noise ratio penalty.

Another problem in the classic dithering model is that the amount of error is difficult to control when the probability density function (PDF) of the dither is uniform ($RPDF$) or gaussian ($GPDF$). Wannamaker, Lipshitz, Vanderkooy and Wright in [2] have proven that dither with a triangular PDF ($TPDF$) can produce an error signal which has a constant variance, as this renders the first and second moments independent of the input. This dither has a larger variance in contrast to $RPDF$ dither and the advantage of this dither is supported with psycho-acoustic tests which show that users prefer a constant noise instead of a non-constant noise. The authors in [2] also prove that, it is not possible to make a classic dither with a constant and lower variance in the quantization error than the one obtained with $TPDF$ dither.

Knowing the advantages and disadvantages of dithering, the aim of this research work was focused on the development of algorithms that measure the need for dithering and to develop a segment dependent dither where dither is added only

to the segments where need dither. Additionally, this research finds an adaptive dither with lower variance than the classic triangular dither.

This thesis is organized as follows. Chapter 2 includes a literature review of quantization, re-quantization, and other techniques to avoid unwanted harmonics like dithering and noise shaping. In Chapter 3, the theoretical foundation of quantization and dithering is derived; providing a basic framework on which Chapter 4 and Chapter 5 are built on.

Chapter 4 is divided in two principal sections. The first section describes methods to measure the need for dithering. The algorithms presented in this section are based in time series hypothesis testing. The second part presents a technique to make a signal dependent dither using constrained optimizations algorithms.

Chapter 5 presents different experiments which show the efficiency of the methods developed in Chapter 4 using synthetic and real audio. Experiments using moving average (MA), autoregressive (AR), and $ARMA$ processes are presented to demonstrate the white noise detection. In addition, experiments using dithering prove the white noise condition of the re-quatization error. Also, in this Chapter, it is presented some experiments to validate the adaptive dithering method. Some optimizations algorithms are tested and contrasted with classical dithering methods.

Finally, the last chapter states conclusion about the results obtained of the different algorithms and discusses future work.

CHAPTER 2
LITERATURE REVIEW

A literature review was done summarizing the most relevant and recent advances in the area of dithering. Most of the previous work in dither addresses methods to avoid unwanted harmonics. This chapter is divided into two parts; the first is about the history of quantization and dithering, and examples of applications that use dithering. The second part describes noise shaping, another similar technique used to avoid harmonics in the spectrum.

2.1 History of Quantization and Dithering

A lot of research has been done in the area of quantization and dithering. Different schemes of analog to digital conversion such as uniform and non-uniform quantization have been developed to sample continuous signals. Details are given in the theoretical framework in Chapter 3. In this research, the quantization and re-quantization are assumed to be uniform where the difference between quantization levels is constant.

The study of quantization properties and its effects increased after the middle of the 20th century. An important mathematical foundation was published by Widrow in 1960 [3] and he recently summarized his work in [4]. Here, he establishes that quantization error can be modeled as a uniform i.i.d sequence under a given set of conditions. He found that the process of obtaining the probability density function (PDF) of the quantized signal is similar to Shannon's Sampling Theorem [5] and named "Area Sampling". The authors explain how, having the relationship

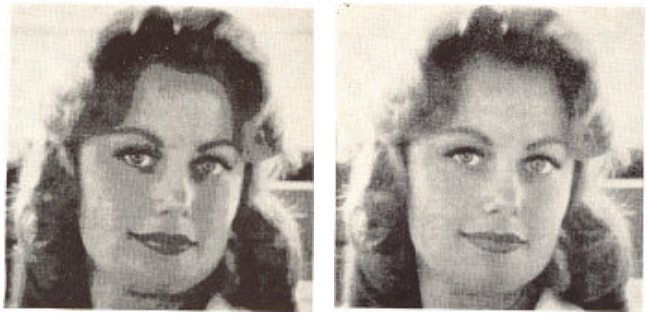
between the input and the output, and the PDF of the output of the quantizer, the original distribution could be recovered. Moreover, Widrow's research work defined the higher order statistics of quantization error and the study of the quantization output moments.

2.1.1 Dithering

The word dithering was originally used during in the Second World War. Aircraft bomb trajectory was more accurate when the airplane was flying, since the vibration reduces the error of moving parts. This vibration was termed dithering. Some initial applications of dithering were introduced by Roberts in his PhD thesis at MIT in [1]. His work was related to the transmission of images in a digital television channel. To transmit a picture, the length of each sample was of at least 6 bits in the Pulse Code Modulation (PCM) standard. This is because the human eye is sensitive to the small changes in intensity. With the use of dithering, Roberts could reduce the resolution to 3 bits. In this work, he added small amounts of noise to the signal before quantization to cause the same effect in the perception of the eye with less bits. Some examples extracted from his thesis are shown in Figure 2.1.

Later in 1964, Schuchman in [6] determined sufficient and necessary conditions of dither to have a minimum loss of statistical properties. For instance, one condition is that statistical dither must be independent of the signal to be quantized. This property will be explained further in Chapter 3.

The state of the art dithering techniques used in industry about dithering were developed at least 20 years ago. Gray and Stokham summarized the most important research about dithered quantizers in [7]. In this work, the theory of the subtractive and non-subtractive dithering and its statistical properties was explained. Also, Gray analyzed the spectra of quantization noise in [8] and summarized principal concepts about quantization, vector quantization, and dithering in [9]. In addition,



(a) Straight PCM 3 bits image

(b) Dithered 3 bits image

(c) Straight PCM 2 bits image

(d) Dithered 2 bits image

Figure 2.1: Examples of dithering made by Roberts in [1]

a strong mathematical foundation for the theory of non-subtractive dithering was developed in the AudioLab at University of Waterloo [10] [11]. These include the first and second order statistics of the system input and output and the introduction of digital dither.

In the middle of the 1980's, digital processing became more widespread and it was necessary to include dither in digital systems. In digital systems, the process of lowering the resolution of a signal is called re-quantization. In [2], a study of digital quantization comes to the same conclusion as with continuous quantization. This work also introduces digital dither. Their work was extended in [12] when triangular PDF dither was used in digital audio. Currently, $TPDF$ is the most popular technique of dithering in one dimensional signals and its benefits were commented

in Chapter 1.

2.1.2 Some Applications of Dithering

Dither is used in various applications. One of the most famous applications was developed in Bell Labs where Jayant and Rabiner used $RPDF$ in speech processing [13]. Another application is presented in [14] where $GPDF$ is used in a high speed digital system for the suppression of the electromagnetic field. Furthermore, in [15], dithering is used in feedback systems to reduce the oscillation at high frequency.

2.2 Noise Shaping

One of the main objectives of this research is to have white re-quantization error when quantizing signals. On the other hand, in some applications it is better to have a non-white noise. For example, human beings have greater perception of sound near to 4kHz. Its possible to modulate (i.e, change the frequency) of the error signal to frequencies where we are less sensitive. In other words, this process changes the shape of the error spectrum to be minimally audible as it is shown in Figure 2.3. This process is known as "noise shaping". In the Figure 2.2, a general scheme for noise shaping is presented, where Q is the quantizer (re-quantizer) and $H(z)$ is a feedback filter. The difference between the input to the quantizer and the output is filtered and added to the input to modulate the total error.

2.3 Summary

In this chapter, a literature review of quantization, re-quantization and dithering was done. In addition, examples of applications that use dithering were presented. Additionally, another technique to avoid unwanted harmonics called noise shaping was presented. The difference between dithering and noise shaping is that the former is added to the signal and produces a total error generally with a white

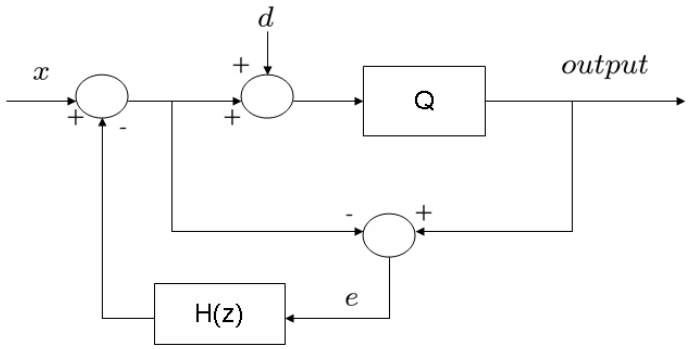
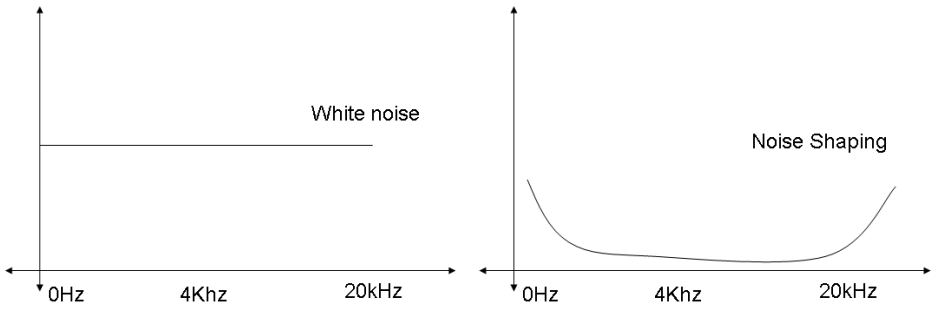


Figure 2.2: General scheme of noise shaping



(a) White noise spectrum (b) Noise shaping spectrum

Figure 2.3: Comparison between white noise and noise shaping

spectrum, and the latter is produced with a feedback system and the spectrum has an arbitrary shape.

CHAPTER 3
THEORETICAL FRAMEWORK

This chapter addresses theory of re-quantization and dithering. The first section defscribes about the statistics of re-quantization noise like mean and variance. Also, the autocorrelation and power spectral density of the quantization noise are presented assuming a set of condition. The second section derives the total error PDF using the convolution theorem and area sampling; it shows that the total error is input dependent for non-subtractive dithering quantization. In addition, the total error statistical moments are defined and later used to obtain the principal properties of triangular PDF dither.

3.1 Re-Quantization

Quantization schemes determine the amplitude resolution of the digital signal. For simplicity, this work assumes a uniform quantization and discrete signals represented in 2s complement binary format. In this case, the amplitude resolution of the digital signal is determined by the number of bits used to represent each sample. The model for the lowered resolution signal is:

$$Q(x[n]) = x_q[n] = x[n] + \varepsilon[n],$$

where Q(x[n]) is the quantization operation and x_q is the lower resolution quantized signal, x the original signal, n indicates the n^{th} sample and ε the quantization noise. The simplest methods of lowering the resolution are truncation, where the

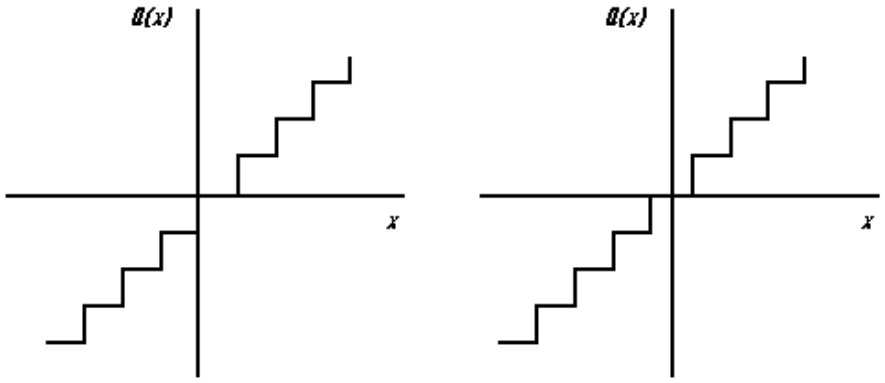
lower significant bits are discarded, and rounding to the nearest integer. Since the numbers are in 2s complement, the truncation operation is

$$x_q[n] = 2^{N-M} \left\lfloor \frac{x[n]}{2^{N-M}} \right\rfloor,$$

where the signal is being truncated from N to M bits, and $\lfloor\ \rfloor$ is the floor operation which rounds to the nearest lower integer. Similarly, the rounding operation is

$$x_q[n] = 2^{N-M} \left\lfloor \frac{x[n]+1}{2^{N-M}} \right\rfloor.$$

Graphical representation of schemes of these quantizers are shown in Figure 3.1. Notice that in each case, the quantized signal is divided and multiplied by the factor 2^{N-M} so that the original and quantized signal have the same amplitude. Let the original signal have quantization levels separated by Δ. The quantized signal then has quantization levels separated by $2^{N-M}\Delta$. For example, if the resolution is being lowered by one bit, the quantized signal has half the resolution and the quantization levels are now separated by 2Δ.



(a) Flooring transfer function (b) Rounding transfer Function

Figure 3.1: Quantization schemes

3.1.1 Statistics of Quantization Noise

If the signal amplitude change is large from sample to sample, then the quantization error is a uniform white noise sequence. In this case, the samples of the quantization noise are uncorrelated between them, and is distributed uniformly from either 0 to $2^{N-M}\Delta$ for truncation, or $-2^{N-M-1}\Delta$ to $2^{N-M-1}\Delta$ for rounding. Under these conditions the quantization error PDF is defined as follows:

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

Given the quantization noise PDF, the statistical moments of a uniform distribution are:

$$E[\varepsilon] = 0,$$
$$E[\varepsilon^2] = \frac{\Delta^2}{12},$$

and in the most general case it is given by:

$$E[\varepsilon^m] = \begin{cases} \frac{1}{m+1}\left(\frac{\Delta}{2}\right)^m, & \text{for } m \text{ even} \\ 0, & \text{for } m \text{ odd} \end{cases} .$$

3.1.1.1 Autocorrelation

The autocorrelation of a signal is the strength of a linear relationship between a pair of points in the signal. The autocorrelation is defined as:

$$a(lag) = E[\varepsilon[n]\varepsilon[n + lag]],$$

where lag is the difference in time (or in samples) between a pair of random variables. When lag is different from zero, assuming an i.i.d zero mean sequence, there linear relationship between the points is zero. On the other hand, when the lag is equal

to zero, then the autocorrelation is equal to the variance of the signal. In summary, the autocorrelation of white noise can be expressed mathematically as:

$$a(lag) = \begin{cases} \sigma^2, & \text{for } 0\ lag \\ 0, & \text{otherwise} \end{cases}.$$

3.1.1.2 Power Spectral Density

Power Spectral Density (PSD) measures the power of a signal in the frequency domain. PSD is defined as the Fourier transform of the autocorrelation. In the case of zero mean white noise it is:

$$P(\omega) = \int_{-\infty}^{\infty} a(\tau)e^{-j\omega\tau}d\tau = \int_{-\infty}^{\infty} \sigma^2\delta(\tau)e^{-j\omega\tau}d\tau = \sigma^2,$$

where $\delta(\tau)$ is the impulse function. This means that white noise has components in all frequencies of the spectrum with the same magnitude.

However, the white noise assumption does not hold for all cases. The quantization process can introduce additional harmonics related to the signal being re-quantized. These added harmonics occur when the quantization noise $\varepsilon[n]$ is highly correlated with signal $x[n]$ and so has harmonic content related to this signal. An example of this is shown in Figure 3.2, where a 1333 Hz sine wave has been re-quantized from 24 to 7 bits. Ideally, the power spectral density of a cosine signal is an impulse at the operation frequency. As can be viewed in the figure, many impulses of small amplitude appear in the spectrum of the re-quantized signal.

3.2 Dither

To avoid these unwanted harmonics, dither is generally added to the signal before quantizing. The purpose of the dither signal is to ensure that the quantization error samples are uncorrelated with the signal being quantized. Ideally, the quantization error signal is independent of the signal being quantized, but it is generally

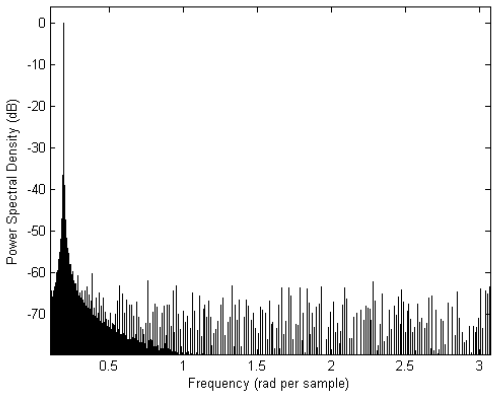
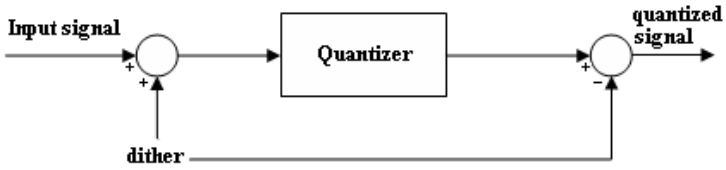


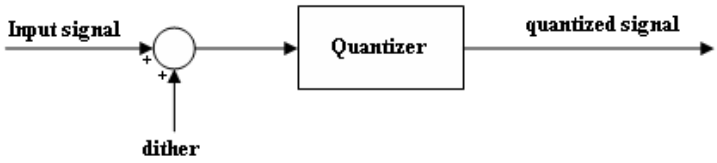
Figure 3.2: Power spectral density of a re-quantized signal

accepted that having uncorrelated first and second moments is sufficient.

Subtractive (SD) and non-subtractive dither (NSD) systems are commonly used. The difference between these schemes is that in the former, dither is subtracted after quantization. General schemes of dithered quantizers are shown in Figure 3.3.



(a) Subtractive dithering



(b) Non subtractive dithering

Figure 3.3: Dithering schemes

The total error (ε) is defined as the difference between input and output signals. Thus the SD total error as:

$$\varepsilon = Q(x+d) - (x+d),$$

where x is the input signal and d is the dither signal. In contrast, the NSD total error is:

$$\varepsilon = Q(x+d) - x, \tag{3.1}$$

This research focuses on *NSD*, so the theoretical framework will be focused on this scheme.

3.2.1 Area Sampling and Dithered Quantization Error PDF

3.2.1.1 Area Sampling

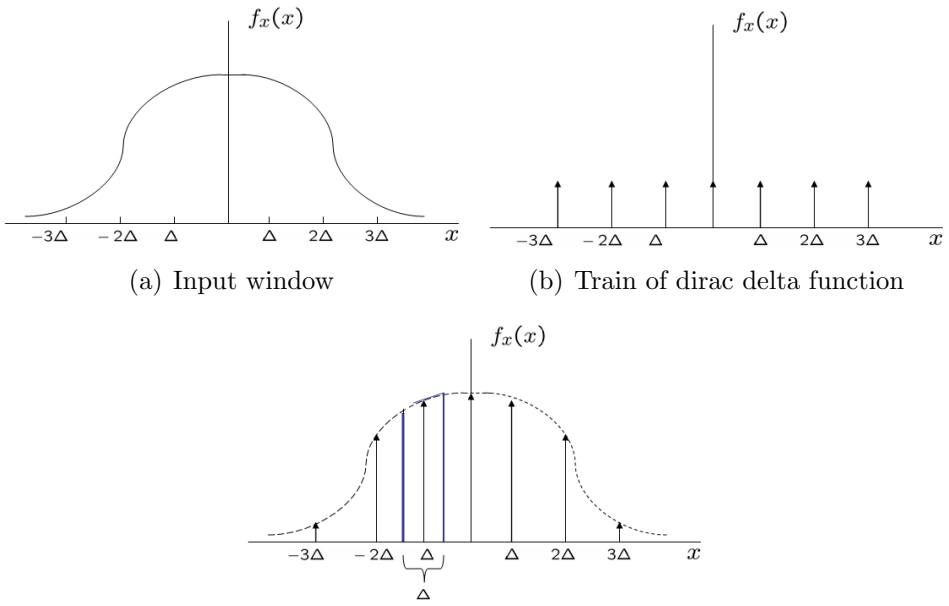
Area sampling is a term used to deduce the PDF of the quantizer error [4]. The quantizer error $q[n]$ (different from total error) is defined as the difference between quantizer input and output. This PDF is a sampled version of the input PDF, with samples taken every Δ. Specifically, the sampling process is the multiplication of the input PDF with a train of impulse functions, then scaled by in the neighborhood of the point in the input PDF. The quantizer error PDF can be represented as follows:

$$p_q(q) = \sum_{k-\infty}^{\infty} \delta(q - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p(y)dy, \tag{3.2}$$

where q is the output signal and $p(y)$ is the input PDF. Figure 3.4 shows the process of obtaining the density of the output. The signal in Figure 3.4(a) is multiplied with Figure 3.4(b) to get the result shown in 3.4(c)

3.2.1.2 Total Error PDF

Knowing the quantizer error PDF, it is possible to deduce the total error PDF using eq. 3.2 . Let $p_{\varepsilon|x}(w,x)$ be the conditional total error PDF in a NSD system.



(a) Input window

(b) Train of dirac delta function

(c) Result PDF

Figure 3.4: Area sampling

Defining $w = x + d$ as the input to the quantizer, the conditional PDF of the quantizer input given x is $p_{w|x}(w, x) = p_d(w - x)$. Using area sampling and Eq. 3.1, the conditional total error is:

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k-\infty}^{\infty} \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_d(w - x)dw. \tag{3.3}$$

Therefore, multiplying by $p_x(x)$ and integrating with respect to x, the marginal PDF of ε is:

$$p_\varepsilon(\varepsilon) = \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x)p_x(x)dx. \tag{3.4}$$

In Eq. 3.4, it is clear that the total error of a NSD is always signal dependent because it is impossible to separate the error ε of the input x.

3.2.2 Total Error Moments

The random variable PDF moments are defined as $E[x^m]$, where m represents the m^{th} moment of such distribution. In many cases, this expression can be calculated using the characteristic function (CF). The CF, by definition, is the Fourier transform of a PDF. The moments are calculated evaluating the CF in the following expression:

$$E[x^m] = \int_{-\infty}^{\infty} x^m f_x(x)dx = \left(\frac{j}{2\pi}\right)^m C_x^{(m)}(0), \tag{3.5}$$

where C_x is the CF of x and $C_x^{(m)}$ denotes the m^{th} derivative of C_x.

In order to find the CF of the quantization error, Eq. 3.3 is redefined as follows:

$$p_{e|x}(e,x) = \sum_{k-\infty}^{\infty} \delta(\varepsilon + x - k\Delta)(v \times p_d)(\varepsilon), \tag{3.6}$$

where "$\times$" denotes convolution and $v(x)$ is a rectangular window which is defined by:

$$v(x) = \begin{cases} 1 & -\Delta < x < \Delta \\ 0 & otherwise \end{cases}.$$

In Eq. 3.1, the total error is the addition (or subtraction) of two independent random variables. Consequently, the PDF of the total error is the convolution of quantized and dithered $PDFs$. Substituting Eq. 3.6 in Eq. 3.4 and using the convolution property, the error PDF is:

$$p_\varepsilon(\varepsilon) = [v \times p_d](\varepsilon)\left[\sum_{k=-\infty}^{\infty} \delta(\varepsilon + x - k\Delta) \times p_x\right](-\varepsilon). \tag{3.7}$$

Taking the Fourier transform of eq. 3.7 and using the convolution theorem, the CF of ε is given by:

$$C_\varepsilon(f) = [sinc(f)C_d(f)] \times \left[\sum_{k=-\infty}^{\infty} \delta\left(-f - \frac{k}{\Delta}\right) C_x(-f)\right], \tag{3.8}$$

$$C_\varepsilon(f) = \sum_{k=-\infty}^{\infty} sinc\left(f - \frac{k}{\Delta}\right) C_d\left(f - \frac{k}{\Delta}\right) C_x\left(-\frac{k}{\Delta}\right). \tag{3.9}$$

Using Eq. 3.5, the m^{th} moment of the error PDF is:

$$E\left[\varepsilon^m\right] = \left(\frac{j}{2\pi}\right) \sum_{k=-\infty}^{\infty} \left(sinc\left(\frac{k}{\Delta}\right) C_d\left(\frac{k}{\Delta}\right)\right)^{(m)} C_x\left(\frac{k}{\Delta}\right). \tag{3.10}$$

Based on this formula, Lipshitz in [16] and [10] demonstrates that if

$$\left(sinc\left(\frac{k}{\Delta}\right) C_d\left(\frac{k}{\Delta}\right)\right)^{(m)} = 0 \quad \forall k \neq 0,$$

then the m^{th} moment is independent of the signal input. Lipshitz also shows that adding uniformly distributed white noise with an amplitude of Δ to a signal uncorrelates the mean of the total error with the input signal. Thus if one uniform noise signal is added, the first moment of the total error is uncorrelated with the input signal. If two are added, then the first two moments are uncorrelated with the input, and similarly for the higher order moments. As mentioned above, it is generally accepted that the first and second moments are the most important, so typically two independent uniform noise signals are added. This gives the $TPDF$ dither used in many audio applications.

3.3 Summary

In this chapter, fundamentals concepts concerning re-quantization and dithering statistics were introduced. The first section was related with mathematic development of re-quantization. In general, the quantization noise is assumed to be a white noise sequence when the amplitude is large from sample to sample. Under this assumption, the quantization noise autocorrelation is an impulse at the first lag, and

its power spectral density is constant for all frequencies. The second section introduced the total error statistics in a non-subtractive dithering scheme. The total error PDF is derived it is shown to is dependent of the input. Also, using the total error statistical moments, the main characteristic of the triangular PDF dither were obtained. In addition, that the first m moments are independent of the input when the dither is the sum of m uniform PDF random variables is shown.

CHAPTER 4
METHODOLOGY

This chapter is concerned with the implementation of new methods of dithering. The first part determines statistical tests for determining the need for dithering. These hypothesis tests are used to measure if the total error is white noise. The second part develops a new adaptive dithering technique that is input dependent and has constant variance. This adaptive dithering is obtained using optimization methods instead of the classical dithering methods that are statistical based.

4.1 Determining the Need for Dither when Re-Quantizing a 1-D Signal

This section deals with statistical methods to measure the need for dither. These methods are time series tests from the field of probability and statistics, introduced for use in quantization by Benitez-Quiroz and Hunt in [17].

4.1.1 Tests in the Time Domain

If the total error ε is white noise, then there will be no added harmonics. The most straightforward method of determining this would simply be testing the serial independence of the signal. It turns out that this is unnecessary, and only the total error needs to be tested. This is because a sufficient condition for not introducing unwanted harmonics during re-quantization is that the total error is an i.i.d. sequence. The reason for this is that an i.i.d. sequence has a white spectrum. As $x_q[n]$ is the sum of $x[n]$ and $\varepsilon[n]$, if $\varepsilon[n]$ is white, then $x_q[n]$ is simply

the original sequence $x[n]$ with added white noise. The tests presented here evaluate the whiteness, or equivalently, the independence of $\varepsilon[n]$ and $\varepsilon[n+k]$ for all $k \neq 0$.

Because $\varepsilon[n]$ is a finite signal, its statistics must be estimated from a finite number of samples. If N samples are used to calculate the estimates, its sample mean and variance can be defined as:

$$\hat{\mu_\varepsilon} = \frac{1}{N}\sum_{n=0}^{N-1}\varepsilon(n)$$

and

$$\hat{\sigma_\varepsilon^2} = \frac{1}{N}\sum_{n=0}^{N-1}(\varepsilon(n) - \hat{\mu_\varepsilon})^2$$

respectively. Similarly, the sample autocorrelation sequence is given by

$$\hat{p_\varepsilon}(k) = \begin{cases} \frac{1}{N}\sum_{n=0}^{N-k-1}\varepsilon(n+k)\varepsilon(n), & 0 \leq k \leq N-1 \\ \hat{p_\varepsilon}(-k), & -(N-1) \leq k \leq 0 \\ 0 & else. \end{cases}$$

If $\varepsilon(n)$ has zero mean, then the sample autocorrelation coefficient can be defined as

$$\hat{r_\varepsilon}(k) = \frac{\hat{p_\varepsilon}}{\hat{\sigma_\varepsilon}^2}.$$

The autocorrelation coefficient at *lag* k gives a measure of the linear dependence between $\varepsilon[n]$ and $\varepsilon[n+k]$. If $\varepsilon[n]$ and $\varepsilon[n+k]$ are independent, then the correlation, and the correlation coefficient will be zero. Thus, a necessary condition for the independence of a sequence is that its sample autocorrelation coefficient be zero for all $k \neq 0$. The tests in the time domain presented here are based on using the sample autocorrelation coefficient to determine if the sequence is independent. Because of the finite number of samples used in the estimation, the sample autocorrelation and autocorrelation coefficient will never be exactly zero. A typical sample

autocorrelation coefficient is shown in Figure 4.1.

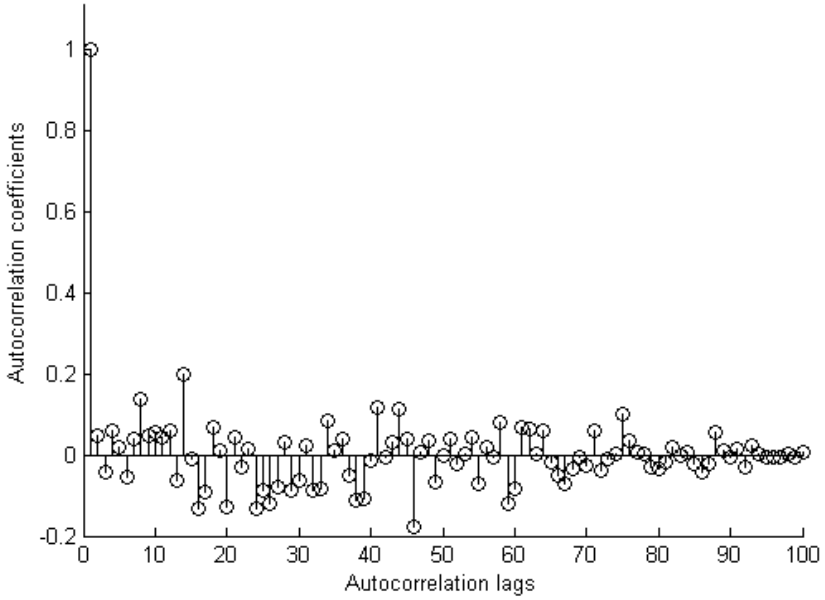


Figure 4.1: Sample white noise's autocorrelation

4.1.1.1 Box-Pierce Test

The Box Pierce Test (Q_{bp}) statistic was introduced in [18], and is used to verify the null hypothesis of white noise. This test uses the first m lags of the autocorrelation coefficient to calculate the following statistic:

$$Q_{bp} = N \sum_{k=1}^{m} r_k^2.$$

Q_{bp} asymptotically approaches a χ^2 distribution with m degrees of freedom for increasing N under the null hypothesis of a white noise signal. In order to make the Q_{bp} statistic closer to its asymptotic distribution, the number of lags must be smaller than the number of samples.

To improve the performance of the test, the error signal is scaled between -1, and 1 and the mean is subtracted. This test is useful to measure when the signal has autoregressive or moving average components.

The Box-Pierce statistic has some problems with the approximation to its asymptotic distribution, and the following test modifies the statistic to achieve a better approximation.

4.1.1.2 Ljung-Box Test

This test is a modification introduced in [19] to the Box-Pierce test. This test has proven more effective when the signal being tested comes from a non-normal distribution. Using the autocorrelation estimator, the statistic is now defined by:

$$Q_{bp} = N(N+2)\sum_{k=1}^{m}\frac{r_k^2}{N-k},$$

where m is the number of lags. Similar to the Box-Pierce test, the distribution asymptotically approaches that of a χ^2 with m degrees of freedom. Also, the performance of the test is improved when N is larger than m.

4.1.2 Test in the Frequency Domain

The tests in the time domain are based on the sample autocorrelation. Similarly, tests in the frequency domain are based on the Fourier transform of the sample autocorrelation, the sample power spectral density. The autocorrelation of a white sequence is an impulse, so its Fourier transform is a constant; it is the variance of the signal:

$$r_\varepsilon(k) = \sigma_\varepsilon^2\,\delta(k) \Leftrightarrow PSD_\varepsilon(\omega) = \sigma_\varepsilon^2.$$

Thus, the test in the frequency domain is used to determine if the PSD of the sequence $\varepsilon[n]$ is a constant. Two modifications to the above equation are made in practice. The sample autocorrelation is used because the signal is finite, and the DFT is used instead of the DTFT for ease of computation. The equation for the PSD estimate, the sample PSD, then becomes

$$PSD_q(l) = \sum_{k=-N+1}^{N-1} r(k)\, e^{-j\frac{2\pi}{N}kl}.$$

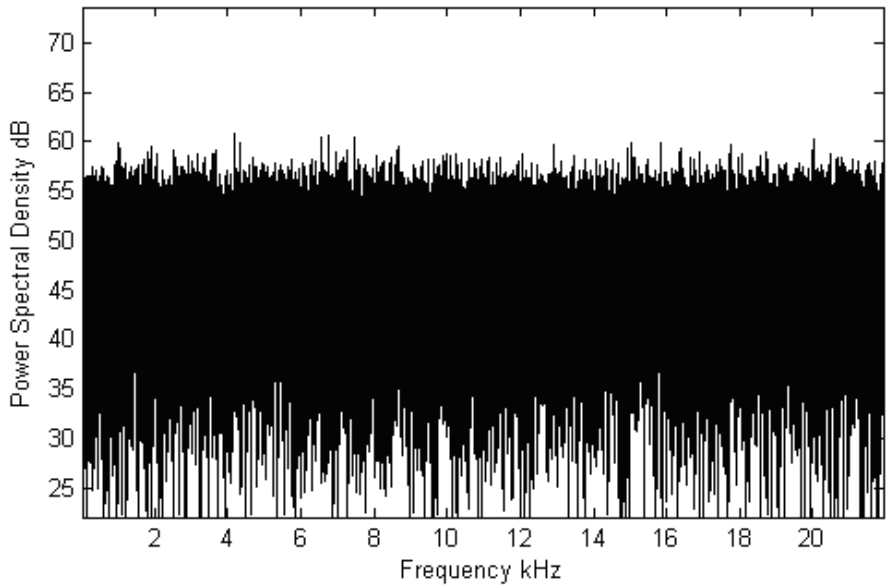


Figure 4.2: Sample power spectral density of the sample white noise.

Figure 4.2 shows the power spectral density of a white noise signal with 1000 samples. As with the autocorrelation, the sample PSD is not constant.

In [20], studies of the spectral response of time series are presented. In this work, the PSD is defined using the data directly instead of its sample autocorrelation. This is given by:

$$A(l) = \sum_{n=-N}^{N-1} \varepsilon[n]\ cos(\tfrac{2\pi\, ln}{N}),$$

$$B(l) = \sum_{n=-N}^{N-1} \varepsilon[n]\ sin(\tfrac{2\pi\, ln}{N}),\ \text{and}$$

$$F(l) = \tfrac{1}{N}(A^2(l) + B^2(l)), -\tfrac{N}{2} < l < \tfrac{N}{2},$$

where F is the PSD estimator. The estimator is the square of the norm of the DFT since the exponential can be defined using sine and cosine functions as follows:

$$e^{\frac{2\pi\ knj}{N}} = cos(\tfrac{2\pi\ knj}{N}) + j\ sin(\tfrac{2\pi\ knj}{N})\ \text{and}$$
$$\left|e^{\frac{2\pi\ knj}{N}}\right|^2 = (cos(\tfrac{2\pi\ knj}{N}))^2 + (sin(\tfrac{2\pi\ knj}{N}))^2.$$

In [20], the following modification to $F(k)$ is suggested.

$$Y(k) = \frac{2F(k)}{\sigma_x^2}.$$

With a large N, $Y(k)$ has a χ_2^2 distribution. If the variance of the sequence is unknown (i.e., the common case), it can be substituted by the sample variance. Knowing the distribution of the signal under the null hypothesis, a goodness of fit test is used to determine if the signal has a χ_2^2 distribution. Some typical goodness of fit tests are the Kolmogorov-Smirnov and the Pearson χ^2 test. The former is used in this work.

4.2 Adaptive Dither

One of the main contributions of this research is a methodology to obtain an adaptive dither which depends on the signal being quantized. This signal dependent dither must have a total error with a constant variance as does $TPDF$ dither. Also, this signal dependent dither must have less variance of the total error, and so less noise, than the typical $TPDF$ dither. To obtain this dither, many statistical

approaches were tried without good results. One of them was proposed by Benitez-Quiroz and Hunt in [21], in which, they linearize the quantizer to deconvolve the PDF using an extended Gaussian Mixture Model approach.

As none of the statistical methods gave satisfactory results, a numerical approach is proposed to find a dither signal using optimization methods. In the following sections, the mathematical problem statement is given and two methods for obtaining a solution are described.

4.2.1 Adaptive Dither Problem Statement

Let the quantized signal be defined as:

$$x_q[n] = Q_r\left(\frac{x[n]}{2^m} + d[n]\right)2^m, \tag{4.1}$$

where $Q_r(x)$ is the rounding operation, and m is the quantization level, $x[n]$ is the input signal, and $d[n]$ the dither signal. The goal is to find a dither $d[n]$ which has a total error with a white spectrum and constant variance. If $x[n]$ is originally an integer, then $\chi = \frac{x[n]}{2^m}$ has an integer and a fractional part. Let the scaled input signal be defined as:

$$\chi[n] = \frac{x[n]}{2^m} = x_i[n] + x_f[n], \tag{4.2}$$

where $x_i[n]$ is the integer part and $x_f[n]$ is the fractional part. The rounding operation is a non-linear process in which the number is approximated to nearest integer. If the input to Q_r has integer parts, they are not affected by the quantizer. This is used to re-write Eq. 4.1 and Eq. 4.2 in Eq. 4.3. $x_f[n]$ is fractional and must be evaluated with the dither $d[n]$ which is unknown so Eq. 4.1 becomes:

$$x_q[n] = Q_r(x_i[n] + x_f[n] + d[n])2^m = (x_i[n] + Q_r(x_f[n] + d[n]))2^m. \tag{4.3}$$

In addition, the scaled quantization error signal is the difference between the input signal and the quantized signal divided by 2^m as shown in the following equation:

$$\epsilon[n] = \frac{x[n] - x_q[n]}{2^m} = \chi[n] - \frac{x_q[n]}{2^m}. \tag{4.4}$$

Replacing Eq. 4.3 in Eq. 4.4 and using Eq. 4.2, the following expression is obtained:

$$\epsilon[n] = \underbrace{x_i[n] + x_f[n]}_{\chi[n]} - \underbrace{x_i[n] - Q_r(x_f[n] + d[n])}_{x_q[n]2^{-m}}, \tag{4.5}$$

$$\epsilon[n] = x_f[n] - Q_r(x_f[n] + d[n]). \tag{4.6}$$

At this point, it is clear that the total error only depends of the fractional part of $\chi[n]$. Hence, the integer part can be ignored in the following sections.

As described in Chapter 3, the autocorrelation of the error is used to measure whether it is white noise. Also, as defined in Chapter 3, the white noise autocorrelation is an impulse at *lag* zero. The desired error autocorrelation for *lags* 0 to $N-1$ is given by:

$$\mathrm{f}(\mathbf{d}) = \begin{bmatrix} f_0(\mathbf{d}) \\ f_1(\mathbf{d}) \\ f_2(\mathbf{d}) \\ f_k(\mathbf{d}) \\ f_{N-1}(\mathbf{d}) \end{bmatrix} = \begin{bmatrix} \epsilon[0]\epsilon[0] & + & \epsilon[1]\epsilon[1] & + & \dots & + & \epsilon[N-1]\epsilon[N-1] \\ 0 & + & \epsilon[0]\epsilon[1] & + & \dots & + & \epsilon[N-2]\epsilon[N-1] \\ 0 & + & 0 & + & \dots & + & \epsilon[N-3]\epsilon[N-1] \\ \vdots & & \ddots & & & & \\ 0 & + & \dots & + & 0 & + & \epsilon[0]\epsilon[N-1] \end{bmatrix} = \begin{bmatrix} N\sigma^2 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \tag{4.7}$$

where $\mathbf{d}$ is the dither vector. Note that the system of equations depends of ϵ and this depends on x_f and $\mathbf{d}$, but x_f is a known constant. Furthermore, the system is clearly non-linear because it has quadratic expressions. Moreover, $\epsilon[n]$ depends on

$Q_r(x)$ which is non-linear. Similar to other non-linear systems, the system in Eq. 4.7 can be solved using an optimization algorithm such as Gauss-Newton or Steepest Descent. To linearize the system, popular optimization algorithms use derivatives or a finite approximation of them, and $Q_r(x)$ is not differentiable, as can be seen in Figure 3.1. Thus, a smooth continuous function is needed for the linearization.

In this research, an approximation is used for the linearization. The simplest estimator of $Q_r(x)$ is to linearize $\hat{Q}_r(x) = x$, so $\frac{d\hat{Q}_r(x)}{dx} = 1$. This approximation shows good results as can be seen in Chapter 5.

4.2.2 Lenvenberg-Marquardt Algorithm for Adaptive Dither

Levenberg-Marquardt (LM) is an iterative optimization algorithm that can be used to find a solution which minimizes a system of non-linear functions. It is popular because it has a double behavior as a Gauss-Newton method and as Gradient Descent. The LM algorithm described below is based on the version published in [22] and [23]. In this work the LM algorithm is used to iteratively obtain the dither vector $\mathbf{d}$. In this algorithm, the update rule to obtain $\mathbf{d}$ is given by:

$$\mathbf{d}^{k+1} = \mathbf{d}^{\mathbf{k}} + \boldsymbol{\alpha}. \tag{4.8}$$

For small values of α and using the first order Taylor expansion to approximate f($\mathbf{d}$) this becomes:

$$\mathrm{f}(\mathbf{d}) = \mathrm{f}(\mathbf{d} + \boldsymbol{\alpha}) - J(\mathbf{d})\boldsymbol{\alpha}, \tag{4.9}$$

where J is the Jacobian matrix. Denoting the iteration error as $\mathbf{e} = \mathrm{f}(\hat{\mathbf{d}}) - \mathrm{f}(\mathbf{d} + \boldsymbol{\alpha})$, where $\hat{\mathbf{d}}$ is the optimum solution, it is desired that the error decreases during each iteration. Hence, it is necessary that $\|\mathbf{e}\|^2 = \mathbf{e}^t\mathbf{e}$ be minimum. Applying Eq. 4.9 to the iteration error gives:

$$\|\mathbf{e}\|^2 = \left\| \mathrm{f}(\hat{\mathbf{d}}) - \mathrm{f}(\mathbf{d}) - J(\mathbf{d})\boldsymbol{\alpha} \right\|^2 . \tag{4.10}$$

Taking the first derivative with respect to $\boldsymbol{\alpha}$ and setting it equal to zero gives Eq. 4.11. This equation is called the normal equation:

$$J^T(\mathbf{d})J(\mathbf{d})\boldsymbol{\alpha} = J^T(\mathbf{d}) \left(\mathrm{f}(\hat{\mathbf{d}}) - \mathrm{f}(\mathbf{d}) \right) . \tag{4.11}$$

Subsequently, $\boldsymbol{\alpha}$ is obtained solving this linear system. Modifying Eq. 4.11, Levenberg and later Marquardt propose a variation to the normal equations which are called the augmented normal equations:

$$N\boldsymbol{\alpha} = J^T(\mathbf{d}) \left(\mathrm{f}(\hat{\mathbf{d}}) - \mathrm{f}(\mathbf{d}) \right) , \tag{4.12}$$

where $N = \mu I + J^T(\mathbf{d})J(\mathbf{d})$. The constant μ is called the damping parameter and it is always positive. The algorithm assumes an initial μ and $\boldsymbol{\alpha}$. If the error is reduced for a given value of $\boldsymbol{\alpha}$, then $\mathbf{d}$ is updated, μ is decreased and a new iteration begins. On the other hand, if the new $\boldsymbol{\alpha}$ increases the error, then μ is increased and $\boldsymbol{\alpha}$ is recalculated. This procedure is repeated until the error is decreased. When the $\mathbf{d}$ is updated the entire process is repeated until the error or the relative change of $\boldsymbol{\alpha}$ are below of a threshold or a maximum number of iterations is reached.

The LM algorithm is said to have a double behavior. This is because when μ is decreased (and it is small) it converges as a Gauss Newton method. In contrast, if μ increases then it converges as a Gradient Descent method.

In the case of dither, it is important that the value of $d[n]$ will be limited in its values to assure that converges to an optimum solution. For this reason, it is necessary to add box constraints (i.e, upper and lower bounds) to the original LM algorithm. Another reason is that the approximations of $Q(x)$ are made in the region between -4Δ and 4Δ. The next section shows a modification to the method

to operate under a set of constraints.

As shown in section 5.2.2, the solution dither vector $\mathbf{d}$ shows high correlation between different audio frames when the system of Eq. 4.7 is solved using LM. The circular autocorrelation estimator is proposed as a method to eliminate this correlation between frames. Let the circular autocorrelation estimator at *lag* L be defined as:

$$f_L(\mathbf{d}) = \sum_{n=0}^{N-1} \varepsilon(\langle n - L \rangle_N)\varepsilon(n), \tag{4.13}$$

where $\langle a \rangle_N$ is the modulus operation. The circular autocorrelation is symmetric at lag $N/2$, so the N equations are not linearly independent causing an error in the optimization libraries. In this case, the system is underdetermined (i.e, less equations than variables) and it is necessary to add more equations to solve it using the same software as before. Assuming that the first frame of length N is analyzed and the total error computed, the next frame is analyzed using $N/2$ samples of the total error calculated and $N/2$ samples of the frame to be analyzed. The system with circular autocorrelation has N samples, $N/2$ unknown variables, $N/2$ know variables and $N/2$ equations. Eq. 4.14 presents the new system of equations for the LM algorithm.

$$\mathrm{f}(\mathbf{d}) = k \begin{bmatrix} \epsilon_{kn}[0]\epsilon_{kn}[0] & + & \epsilon_{kn}[1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-2]\epsilon_u[N-2] \\ \epsilon_u[N-1]\epsilon_{kn}[0] & + & \epsilon_{kn}[0]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-2]\epsilon_u[N-1] \\ \epsilon_u[N-2]\epsilon_{kn}[0] & + & \epsilon_u[N-1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-3]\epsilon_u[N-1] \\ \vdots & & \ddots & & & & \\ \epsilon_u[\frac{N}{2}]\epsilon_{kn}[0] & + & \epsilon_u[\frac{N}{2}+1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_{kn}[\frac{N}{2}]\epsilon_u[N-1] \end{bmatrix} = k \begin{bmatrix} N\sigma^2 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \tag{4.14}$$

where ϵ_u and ϵ_{kn} are the unknown and known error samples respectively.

4.2.2.1 Box Constrained Levenberg-Marquardt Algorithm

This algorithm was proposed by Kanzow, Yamashita and Fukushima in [25]. The algorithm is called a projected Levenberg-Marquardt (PLM) method because $\mathbf{d}$ is a projection onto the desired constrained space. This space is a set of upper and lower bounds for the vector $\mathbf{d}$.

The update rule in PLM is defined as the projection onto the desired region. The update sequence is given by:

$$\mathbf{d}^{k+1} = P_X\left(\mathbf{d}^k + \boldsymbol{\alpha}\right), \tag{4.15}$$

where P_X is the projection operation. Similar to the unconstrained LM algorithm, if the error is reduced, then μ is reduced and a new iteration begins. Otherwise, the updated value is defined as $\mathbf{d} = P_x\left(\mathbf{d}^k - t_k J^T(\mathbf{d})\right)$, where t_k must be well chosen [25] to have an error in decreasing direction.

4.2.3 Spectral Projected Gradient Optimization

Spectral projected gradient (SPG) is an algorithm which minimizes an objective function in a closed region, in this method denoted a box. The method is a modification to the Barzain-Borwein gradient descent algorithm. The algorithm was introduced by Birgin, Martinez and Raydan in [26] and modified by themselves in [27]. The method has the following objective function:

$$F(\mathbf{d}) = \frac{1}{2}\sum_{i=0}^{N-1}\left(f_i(\mathbf{d})\right)^2 = \frac{1}{2}\left\|\mathrm{f}(\mathbf{d})\right\|^2 = \frac{1}{2}\mathrm{f}(\mathbf{d})^T\mathrm{f}(\mathbf{d}). \tag{4.16}$$

Similar to LM, the spectral projected gradient finds a vector $\mathbf{d}$ which minimizes the objective function. The update rule to obtain $\mathbf{d}$ is given by:

$$\mathbf{d}^{k+1} = \mathbf{d}^k + \zeta\mathbf{p}^k, \tag{4.17}$$

where ζ is the step size and $\mathbf{p}^k$ is the search direction defined as:

$$\mathbf{p}^k = P_x\left(\mathbf{d}^k - \frac{1}{\kappa_k}\nabla F(\mathbf{d}^k)\right) - \mathbf{d}. \tag{4.18}$$

The parameter κ^k is obtained using the Barzain-Borwein gradient descent method described in [28] where the parameter is estimated using the slack variables, $s^k = x^{k+1} - x^k$ and $y^k = \nabla F(\mathbf{d}^{k+1}) - \nabla F(\mathbf{d}^k)$. Therefore, the factor is given by:

$$\kappa^{k+1} = \frac{(\mathbf{s}^k)^T\mathbf{y}^k}{(\mathbf{s}^k)^T\mathbf{s}^k}. \tag{4.19}$$

Note that κ can take any positive value. To avoid very large or very small values of κ, the method inserts an upper and lower bound. Therefore, the parameter becomes:

$$\kappa^{k+1} = min\left\{\kappa_{max}, max\left\{\kappa_{min}, \frac{(\mathbf{s}^k)^T\mathbf{y}^k}{(\mathbf{s}^k)^T\mathbf{s}^k}\right\}\right\}. \tag{4.20}$$

Finally, the vector ζ is calculated using a non-monotone line search [26]. In this line search, ζ_k is found if $F(\mathbf{d}^{k+1}) \leq F_{max} + \gamma\zeta_k\nabla F(\mathbf{d}^k)^T\mathbf{p}_k$, where $\gamma \in [0, 1]$. Otherwise, the value is iteratively obtained using the following update rule:

$$\zeta_{tmp} = -\frac{1}{2}\frac{\zeta^2\nabla F(\mathbf{d}^k)^T\mathbf{p}_k}{F(\mathbf{d}^{k+1}) - F(\mathbf{d}^k)}.$$

If ζ_{tmp} is between the upper and the lower bound of κ then $\zeta^k = \zeta_{tmp}$, otherwise, $\zeta = \zeta/2$.

4.3 Summary

The hypothesis test methods presented were Box-Pierce, Ljung-Box and a frequency test. The second section presented a new technique called adaptive dithering. Adaptive dithering uses Projected Levenberg Marquardt and Spectral Projected Gradient optimization algorithms to solve a non-linear system of equations. This technique has constant, user defined variance of the total error.

CHAPTER 5
EXPERIMENTS AND RESULTS

This chapter shows experiments about the need for dither and adaptive dithering in one dimensional signals. In the first part, there are experiments with synthetic and real audio to determine if the total error is white noise. Also, in the first section, examples of the effectiveness of the methods when the data comes from AR, MA and $ARMA$ process are shown. The second part of this chapter deals with adaptive dither in synthetic and real audio. The experiments are designed to test if adaptive dithering reaches the desired variance in the total error at different levels of quantization and at different values in the desired variance.

5.1 Experiments for Measuring the Need for Dither

The purpose of these experiments is to test the algorithms described in sections 4.1.1 and 4.1.2 which measure the need for dither. The experiments are statistical based hypothesis testing. This type of test has two types of error: saying the null hypothesis is false when it is true (type I error) and saying it is true when it is false (type II error). The tests are typically designed to have a user defined type I error. This methodology is used here and the type I error has been set at 10%. The null hypothesis in these experiments is that the signal is white noise. Thus, the error level has been selected so that 10% of the time the signal will be said to be not white when it really is.

5.1.1 Experiments using an AR Process

The experiments in this and the next two sections are designed to test the relative performance of the different tests with synthetic data. The purpose of this section is to measure the performance of the hypothesis tests when the data comes from an autoregressive process as shown in Figure 5.1. The term $z^{-}1$ in the figure is a delay in time domain, so, the linear difference equation is given by $o[n] = x[n] + ao[n-1]$, where o is the output of the AR process.

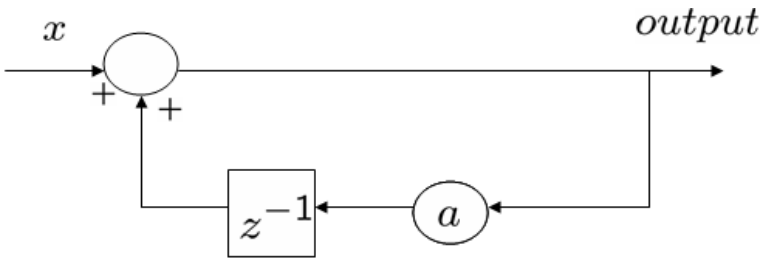


Figure 5.1: Autoregressive process scheme.

The level of the constant a controls the relative level of whiteness of the output signal. The input signal is a zero mean white noise uniformly distributed with 10000 samples. The two time domain tests and the frequency domain test of Chapter 4 are used to test the output signal as the constant a is varied from 0 to 1. As can be seen from Figure 5.2, when the constant is set to zero the output is white noise, and as the constant increases the relative whiteness of the output signal decreases. The three tests reject the null hypothesis of white noise for relatively small values of a.

The probability value (p-value) in Figure 5.2 can be interpreted as the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone under the null hypothesis. In this case, the three tests are fairly close, with the Ljung-Box test correctly identifying the signal as being non-white when the magnitude of a is only about 0.2.

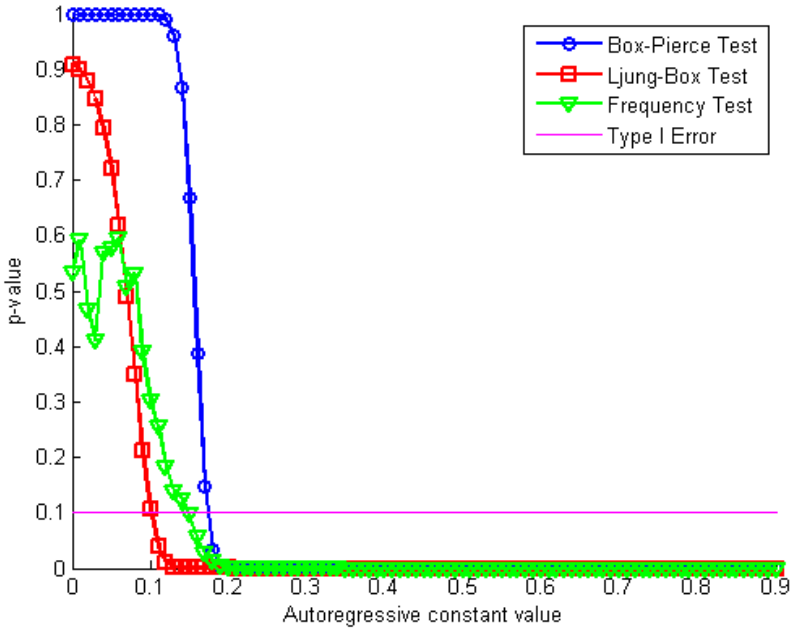


Figure 5.2: P-values of the different test and the p-value threshold.

5.1.2 Experiments using an MA Process

The purpose of this experiment is to evaluate at what level of whiteness the tests reject the null hypothesis when the data comes from a moving average process. This experiment uses synthetic data from a first order moving average process as shown in the Figure 5.3. In this case, the linear difference equation is given by $o[n] = x[n] + ax[n-1]$

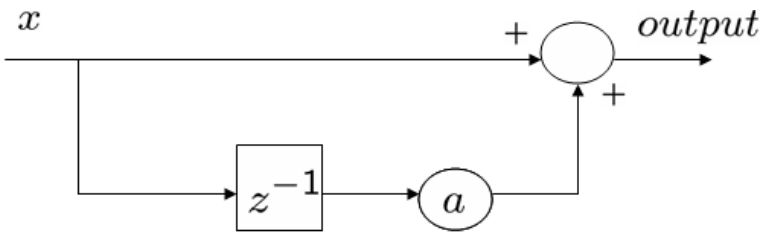


Figure 5.3: Moving average process scheme.

Similar to the AR process, the constant a controls the level of whiteness of the MA process. The input signal is zero mean white noise with 10000 samples. As seen in Figure 5.4, when the constant a is close to zero the white noise tests does not reject the null hypothesis. On the other hand, when the constant is increased the white noise tests reject the null hypothesis. The Ljung box test is the best test because it rejects the null hypothesis for a smaller value of a than the other tests

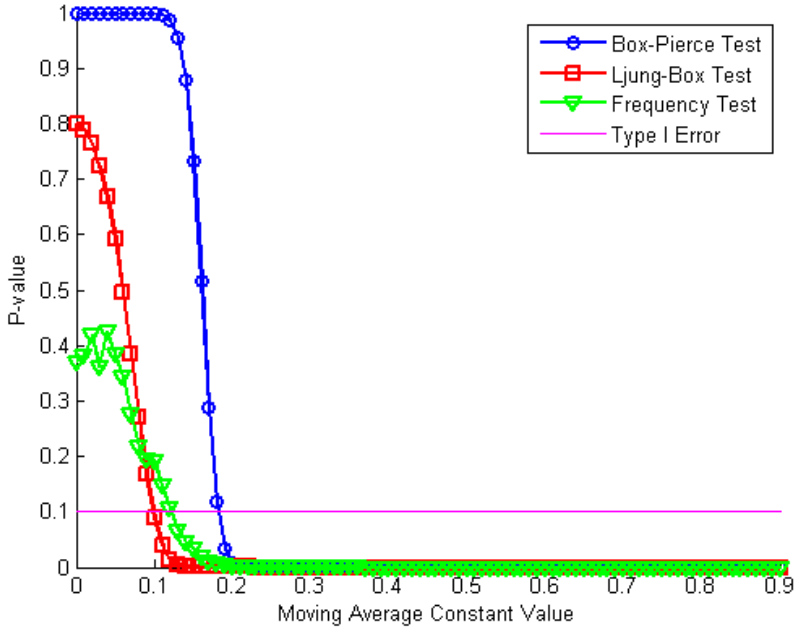


Figure 5.4: P-values of the different test and the p-value threshold.

5.1.3 Experiments using a ARMA Process

Analogous to the MA and AR process, the goal of this experiment is to evaluate the different tests when the data comes from a first order $ARMA$ process. The input signal has the same characteristics of the last section. The process generation is shown in Figure 5.5. Note that the $ARMA$ process is the concatenation of an AR process and an MA process, so, its linear difference equation is given by $o[n] = x[n] + ax[n-1] + bo[n-1]$.

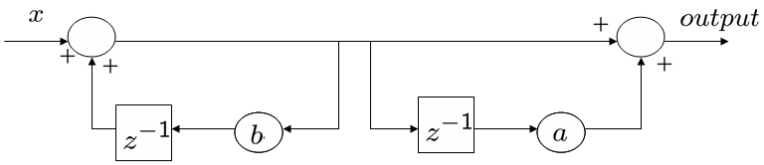
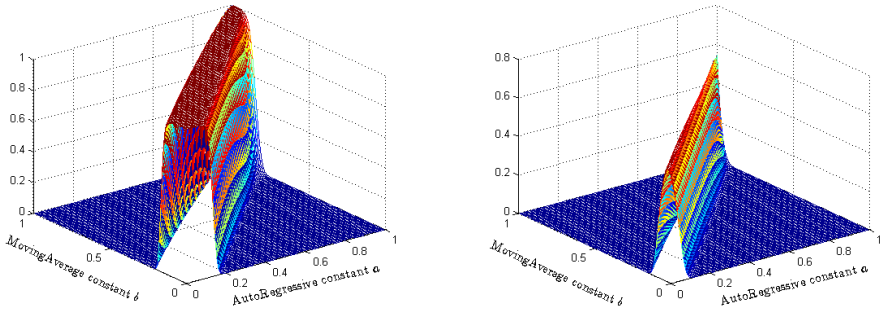


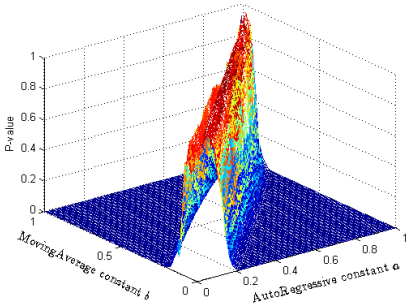
Figure 5.5: ARMA process scheme.

In this case, the signal is controlled using an autoregressive constant a and the moving average constant b. The results of varying the constants a and b between 0 and 1 are shown in Figure 5.6.



(a) Box Pierce test

(b) Ljung Box test



(c) Frequency test

Figure 5.6: White noise test in ARMA signal

When the constants are near zero, the output is white noise, and as the constants increase, the relative whiteness of the output signal decreases. When the constants a and b are the same, the signal is white noise. This is because the pole and the zero of the transfer function cancel when they are the same. The Ljung Box

test is better than the other tests because it rejects the null hypothesis at smaller values of a and b than the Box-Pierce and Frequency Test.

5.1.4 Audio Experiments

5.1.4.1 Experiments using Synthetic Audio

This set of experiments use a full scale 1333Hz cosine wave with 24 bit precision. Each segment has a length of 10000 points and 1250 lags are used in the autocorrelation tests.

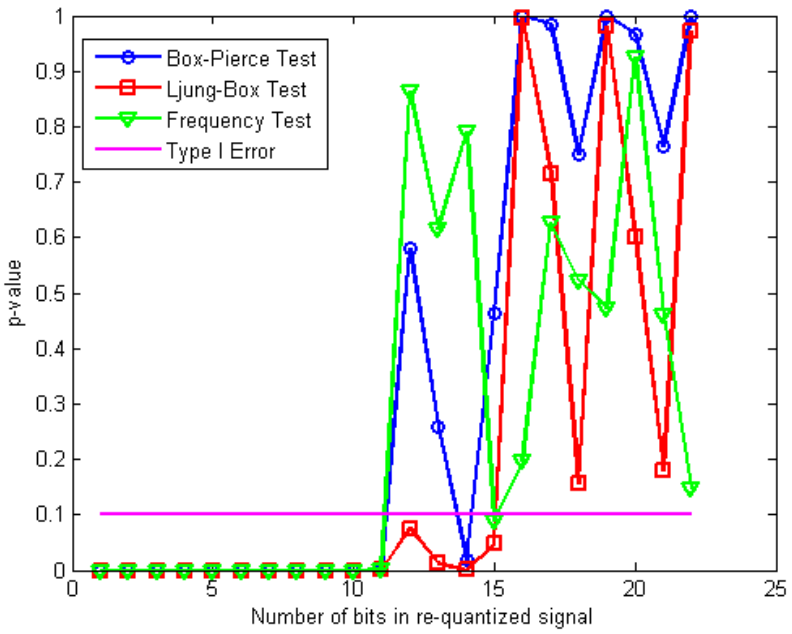


Figure 5.7: P-values of different levels of quantization for a sine signal

This first experiment is done to determine at what bit level the quantization noise ceases to be white. Here, the signal is re-quantized to different bit levels varying from 22 down to 1 bit.

As can be seen in Figure 5.7, the tests have p-values above 0.1 until the signal is quantized around to 16 bits. At this point the p-values of the tests begin to decrease, each one showing that the total error is no longer white when the signal

is re-quantized to values between 12 and 14 bits. Below to 12 bits the hypothesis test has p-values near to zero. The Box-Pierce test is the most monotonic, with the Frequency test having the most ripples near the hypothesis rejection level meaning less stability in its rejection.

Another experiment was performed using the same signal as before. The purpose was to observe the unwanted harmonics when the amplitude of the signal is changed. This experiment using synthetic audio uses a 1333Hz cosine wave with 24 bit precision, and re-quantizes to either 16 or 12 bits. In this experiment, the amplitude of the cosine is decreased until unwanted harmonics appear in the re-quantized signal. It was found that the total error is classified as non-white before the harmonics are visible in the spectrum. When re-quantizing to 16 bits, the tests reject the null hypothesis when the cosine has been reduced to -20 dB or more. However, the harmonics are not visible in the PSD until the signal has been reduced by at least -20.9 dB of its maximum value. The results are similar when re-quantizing to 12 bits. The tests reject when the amplitude has been reduced by at least -10dB, but visible harmonics appear when the amplitude has been reduced by at least to -13 dB.

5.1.4.2 Experiments using Real audio

The final experiments of this section used real audio with 24 bit precision and a 44.1 kHz sampling rate. Since an audio signal can change over time, the procedure is to segment the signal, and apply the tests on each segment. This is done to strengthen the stationary assumption. As in the previous test, each segment is 10000 samples in length, and the autocorrelation tests used a lag of 1250. The purpose of this experiment is to determine the number of frames that are rejected at different levels of quantization. This is done because it is assumed that an increase in the resolution produces a decrease in the number of rejected frames. For this experiment the audio signal has 711 frames.

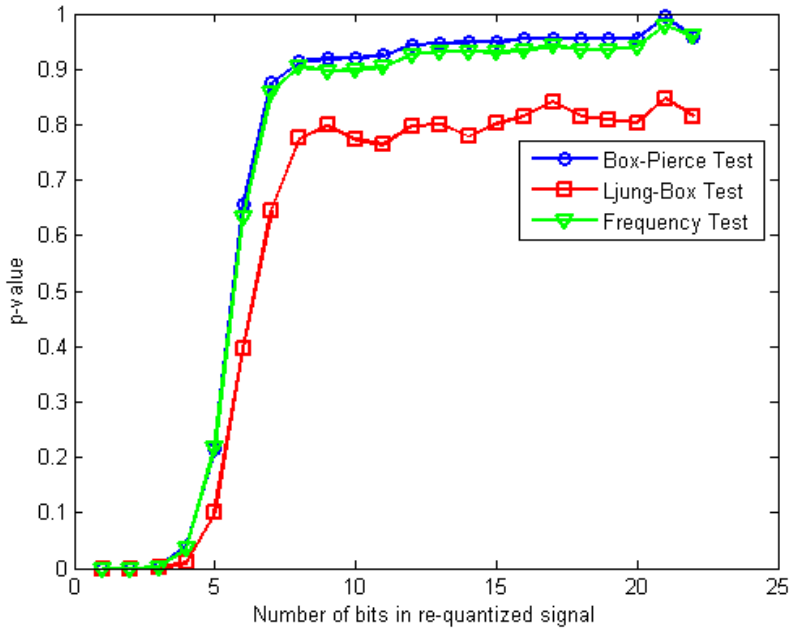


Figure 5.8: Percentage of frames accepted as having white quantization noise

Figure 5.8 shows the percentage of frames that were accepted when the re-quantized signal has q bits of precision. As can be seen in the figure, the general tendency is that re-quantizing to a large number of bits has a white noise total error, and the opposite for re-quantizing to a small number of bits. As can be expected, there are frames that were accepted by one test and rejected by another.

A second experiment was performed in real audio to know at what bit depth in the re-quantized signal the tests do not reject the null hypothesis. In addition to looking at all the frames together, a study of only one frame was made. This frame was selected randomly and the same tests as above were performed for different re-quantization levels. The results are shown in Figure 5.9.

As can be seen in the figure, the Box-Pierce test is again the most monotonic, with the frequency test having the most ripples. This suggests that the frequency test is less monotonic rejecting the null hypothesis than the autocorrelation tests.

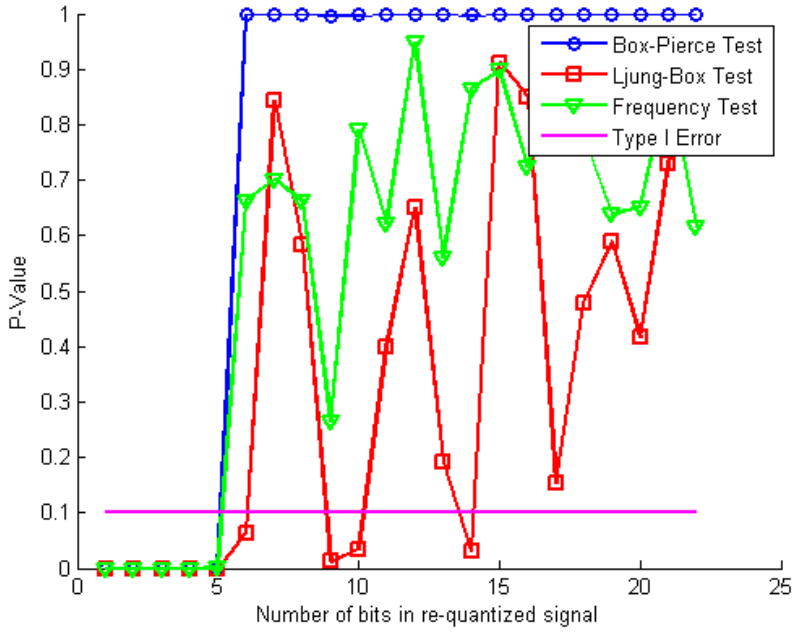


Figure 5.9: P-value of the one frame at different quantization level

5.1.4.3 Real Audio with Additive Dithering

If the tests proposed in Chapter 4 are used to measure white noise in the total error, then dither does not have to be added to the whole signal, only to the segments that do not pass the tests. As seen in the previous section, re-quantizing a segment to 4 bit resolution introduces unwanted harmonics and so does not have white total error. To see the effectiveness of adding dither, and if the test will show that the signal has been effectively dithered, the tests were run after adding dither. The results are shown in Figure 5.10.

Figure 5.10 shows that the dithered segment of real data has white total error. In addition, the total error is not rejected by the white noise hypothesis test, suggesting that dithering effectively uncorrelates the samples from the quantization total error.

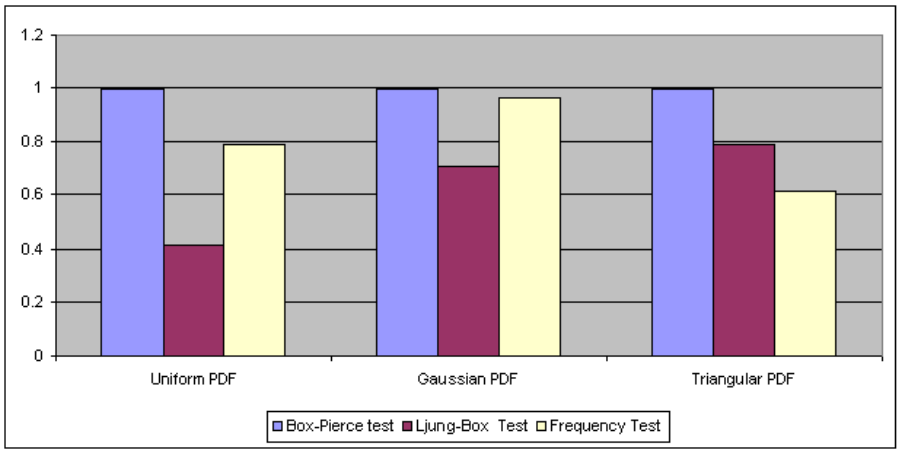


Figure 5.10: Variance for different types of dithers.

5.1.5 An Application to Measure the Need for Dither

One of the goals of this thesis was to implement the algorithms of chapter 4 in Ansi C/C++. This is done beacuse the compiled C code is faster than Matlab^{TM}and other scripting languages. The program was written in Visual C++ and uses the GNU Scientific Library (*GSL*) for general mathematical functions. Other functions have been programed in order to have the same functionality of Matlab^{TM}.

The program processes a standard 24 bit audio file in WAV format. The program segments the input wave file into frames of 10000 samples. After that, it truncates each sample to 16 bits using a shift right operation. It then calculates the p-value of the Ljung-Box, Box-Pierce and frequency test. When any of the tests rejects the null hypothesis, dither is added to this frame. To avoid the abrupt change of noise in the signal, the program increases and decrease the noise level gradually.

5.1.5.1 Specifications

The Figure 5.11 shows the application window. The check box on the right of the figure indicates the possible outputs of the program. These include the dithered wave file and the output text file that writes the frames that do not pass the tests.

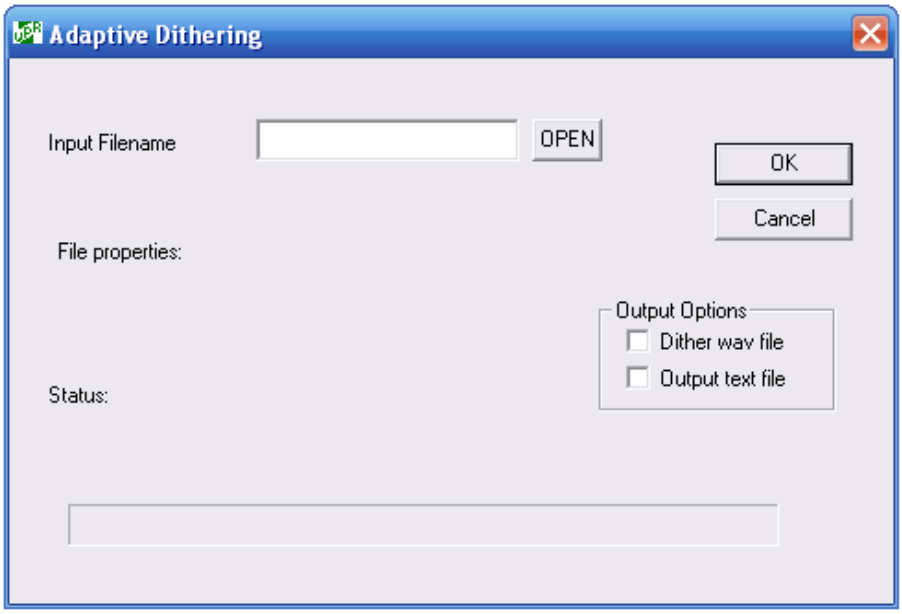


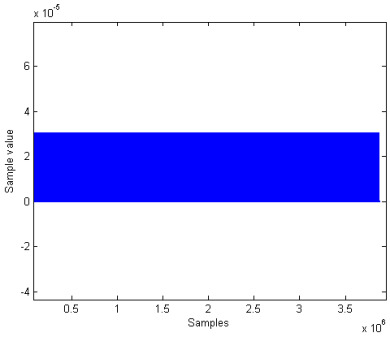
Figure 5.11: Application to measures dither

In the file properties edit box, the program writes information about the sampling frequency, the number of samples and the number of channels (mono or stereo). Meanwhile, in the status edit box, the program indicates if the signal is being analyzed or the output files has been written.

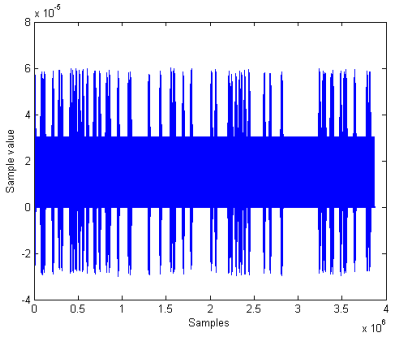
5.1.5.2 Testing the Segment Dependent Dither C Program

The following experiment evaluates if dither is added in the frames where the test rejects the white noise hypothesis. This experiment was done using real audio data. Figures 5.12(b) and 5.12(a) show the total error signal with segment depended dithering and undithered quantization respectively. Figure 5.12(c) shows the difference between these signals.

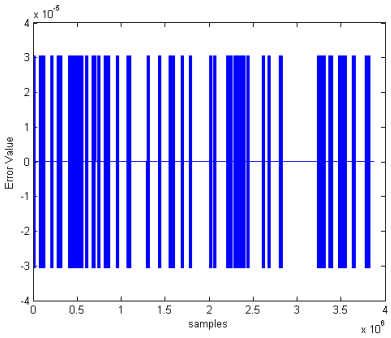
In Figure 5.12(c), it can be seen that when the total error has long periods of zero, meaning that in these sectors dither was not added. Figure 5.12(b) shows that dither is added in some segments of the audio signal in contrast to Figure 5.12(a) where the total error is uniform.



(a) Total error signal without dither



(b) Total error signal with adaptive dither



(c) Difference between Quantized Signal with and without Adaptive Dither

Figure 5.12: Comparison between an a undithered quantized signal and SDD quantized signal

The following experiment is a comparison of the sample variance of the total error when the signal of Figure 5.12(a) is quantized with different types of dither.

Table 5.1 shows the variances of the total error in non-subtractive dithering scheme. As can be shown, the quantized signal with triangular PDF dither in the frames that do not pass the tests has the lowest variance.

Now that the program has been shown to be working correctly, the total error is analyzed. Figure 5.13 illustrates how the total error increases and decreases gradually avoiding abrupt changes of noise in the total error. Also note that in the frames which do not need dither the noise is apparently uniform i.i.d. sequence.

Dither	Variance
Uniform PDF	0.1666
Gaussian PDF	0.1734
Triangular PDF	0.2502
Frame dependent triangular PDF	0.1182

Table 5.1: Variance for different types of dithers.

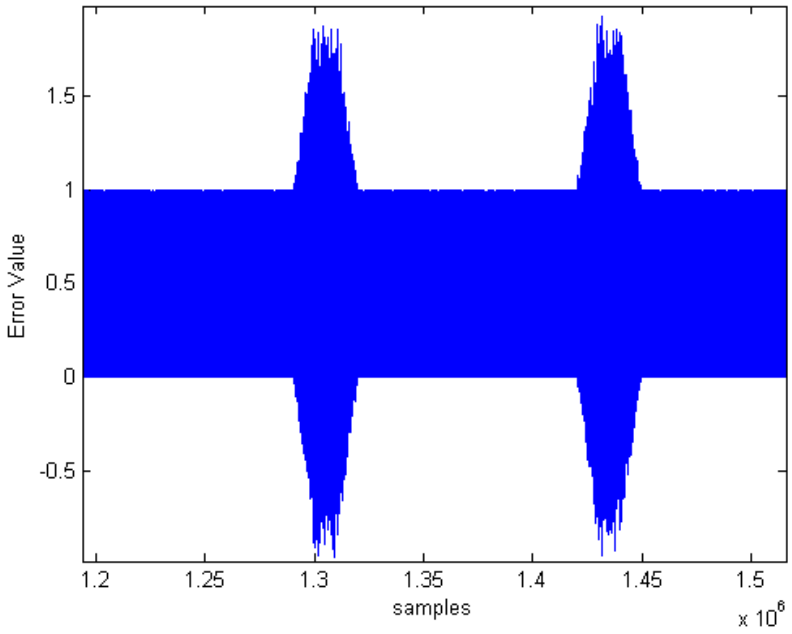


Figure 5.13: Total error in a SDD

5.2 Adaptive Dither

5.2.1 Adaptive Dither in Synthetic Audio

This set of experiments use a 10% of full scale 1333Hz cosine wave with 24 bit precision.

This experiment seeks to measure the difference between the total error variance obtained with Levenberg-Marquardt algorithm (LM) and Spectral Projected Gradient(SPG) and the desired variance. The total error variance of adaptive

dither techniques described in Chapter 4 are compared with the variance of classical dither techniques. The input signal is dithered with adaptive dither, uniform PDF dither($RPDF$), triangular PDF dither ($TPDF$), and Gaussian PDF dither ($GPDF$) and re-quantized to 16 bits. The desired variance of the total error for LM and SPG has been set to 0.150 as this is below the mean variance of the total error when dither has a $RPDF$ (i.e, near 0.17) or a $TPDF$ (i.e, near 0.250). The length of the frames has been set between 1000 and 10000 samples.

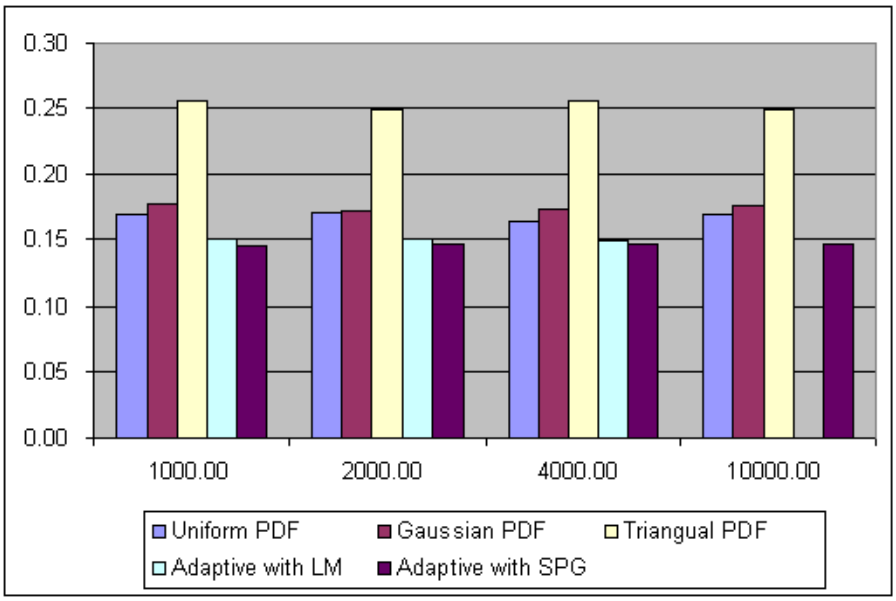


Figure 5.14: Total error variance at 16 bits in re-quantized signal

Figure 5.14 shows that the absolute error between the desired and the sample variance for the adaptive dither is around 0.01. In the case of Gaussian PDF and uniform PDF dithers the variance can change depending of the input signal in contrast with triangular PDF dither, which it is independent of the signal. In the case of the adaptive dithering with LM and SPG, the desired variance was approximately reached.

The previous experiment measures the variance of the total error when the re-quantized signal has 16 bits and it is necessary to measure if adaptive dither works at

different bit depth in the re-quantized signal. The goal of the following experiment is to measure adaptive dither variance at different levels of quantization. The signal is quantized with adaptive dither and compared with classic techniques. The input signal has 1000 samples, the variance for adaptive dithering is set at 0.150 and the quantization levels q are 8, 13, 16 and 19.

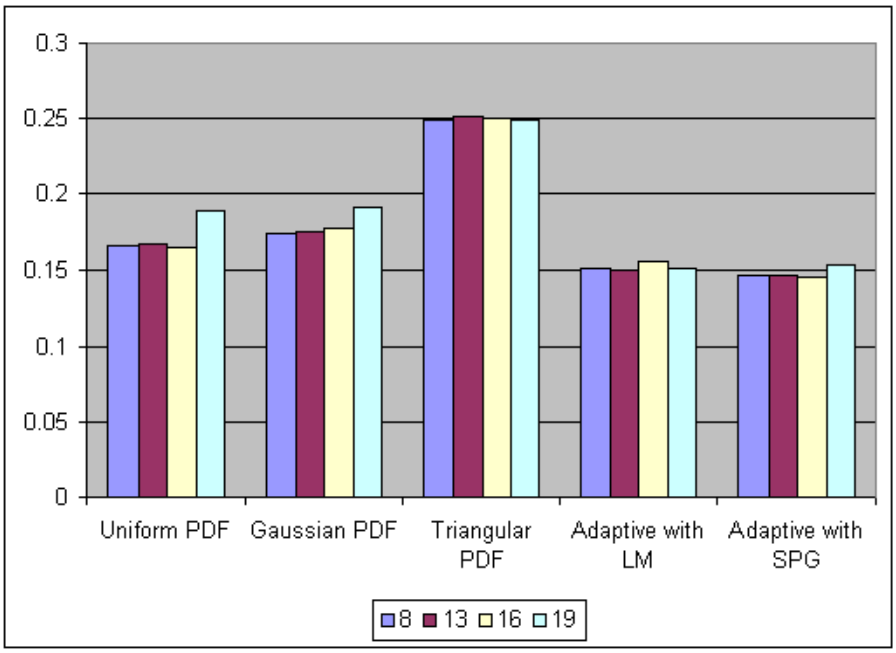


Figure 5.15: Variances of the total error at different levels of quantization.

Figure 5.15 presents the variance of the total error at different levels of quantization. The results show that the adaptive dither noise achieves the desired variance with a small error margin. Also, $ANSD$ with LM is more accurate for this data than the adaptive dithering with SPG. Furthermore, Figure 5.15 shows that for uniform PDF and Gaussian PDF the total error variance change depending of the quantization level.

The aim of the following experiment with synthetic audio is to evaluate if the desired variance of the total error is reached when this is set at different values. Specifically, this desired variance is varied between 0.11 and 0.3.

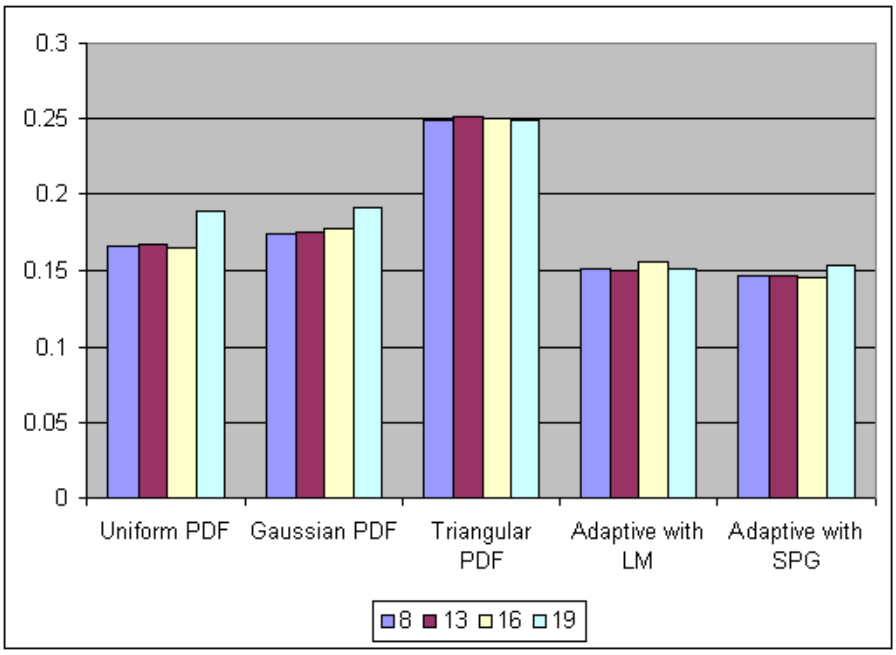


Figure 5.16: Variances of the total error with different desired variance.

Figure 5.16 shows that the sample variance of the total error is near to the desired variance. Different than the previous experiment in table 5.15, SPG presents better results than LM.

5.2.2 Experiments with Adaptive Dither and Real Audio

The purpose of the following experiments is to test the adaptive dithering using real audio under various conditions. In these experiments dither is applied to each frame before to be re-quantized. This algorithm was implemented in C++ using LEVMAR [29] for LM and Ool Optimization Package [30] for SPG. The experiments used real audio with 24 bit precision and a 44.1 kHz sampling rate. The optimization packages speed is related with the number of variables, so the program segments the data into frames. Furthermore, to analyze the signal, the program needs the number of bits, the filename, the length of the frame and the number of frames to analyze.

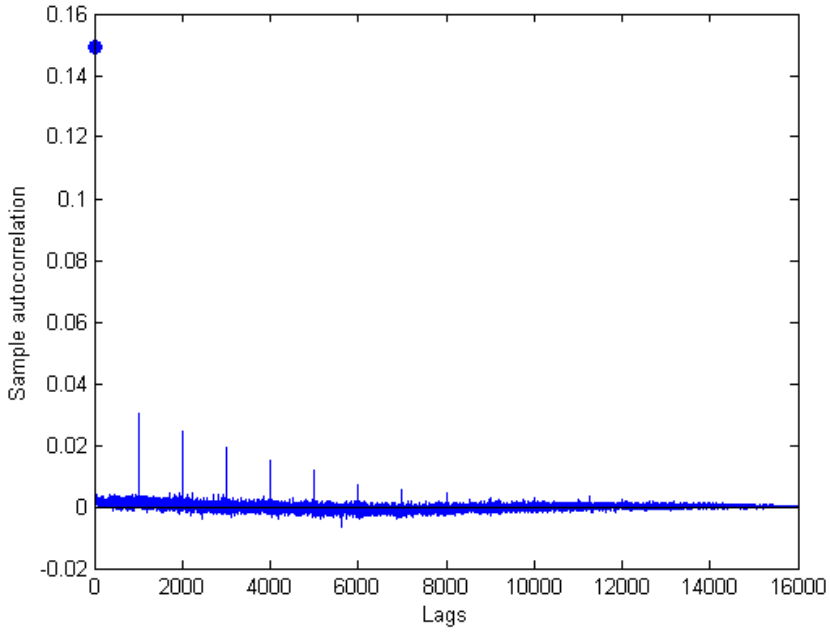


Figure 5.17: Sample autocorrelation of $ANSD$ with LM

As the frames are relatively short, it was necessary to view the autocorrelation of the entire dataset to test if the total error signal was white noise. As seen in Figure 5.17, in the case of $ANSD$ with LM the total error signal has harmonics resulting in a bad quantization. The Ljung Box test was used and it rejected the null hypothesis of white noise.

This occurs because the first and the last error signals are contributing more to the optimization than the other samples in Eq. 4.7. Similarly, the second most influential samples are ε_2 and ε_{N-1} and so on. This indicates that the error samples from the middle has a smaller effect than the others, showing a higher relationship between the total error frames as shown in Figure 5.18. The only way found to correct this side effect was using the circular autocorrelation estimator and redefining the system of non-linear equations as can be shown in Eq. 4.13 and Eq. 4.14 respectively.

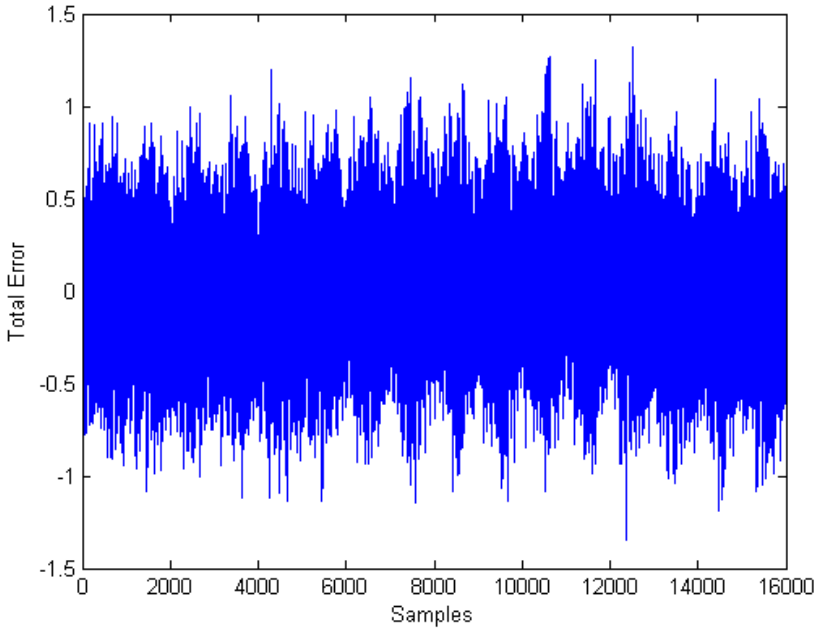
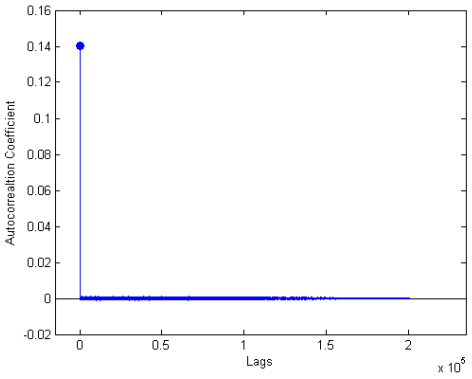
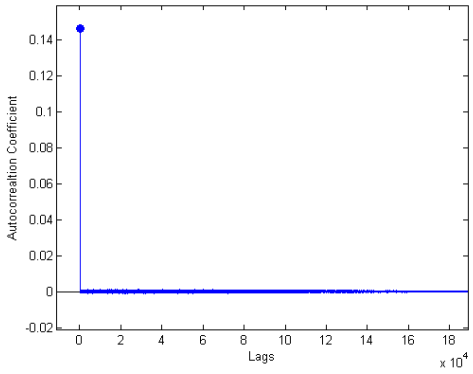


Figure 5.18: Sample autocorrelation of $ANSD$ with LM

The following experiments with adaptive dithering and real audio are done to show that the methods reach the desired variance in the total error when the input is segmented in frames. These experiments use a set of 200000 samples segmented into frames with lengths of 1000, 2000, and 4000 samples for LM. For SPG the lengths of the frames were 1000, 2000, 4000, and 10000 samples. The desired variance in the total error is set at 0.150. The audio file has been written at 24 bits and the re-quantized signal at 16 bits. Figure 5.19 shows the sample autocorrelation for $ANSD$ with a frame length of 1000 samples. As seen in this figure, the autocorrelation is white noise for LM and SPG. Table 5.2 shows that the total error in an adaptive dithering does not reject the null hypothesis of white noise for Ljung-Box Test, Box Pierce and Frequency tests. Also, Figures 5.19, 5.20, 5.21 and 5.22 show the sample autocorrelation for frame lengths of 2000 and 4000 and 10000 samples. For these, the same behavior can be observed as in Figure 5.19. In the same way, tables 5.3, 5.4 and 5.5 show that the set of samples are not rejected by the hypothesis tests.



(a) Autocorrelation for ANSD with LM

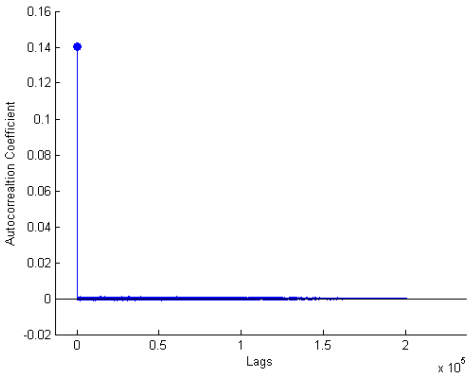


(b) Autocorrelation for ANSD with SPG

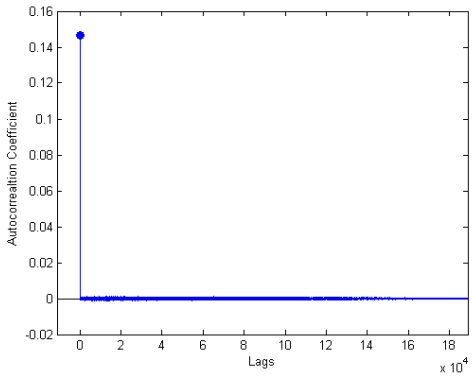
Figure 5.19: Autocorrelation of 200000 lags with frames of 1000 samples

	Ljung Box	Box Pierce	Frequency
Adaptive dither with LM	0.9562	1.0000	0.3905
Adaptive dither with SPG	0.8287	1.000	0.3463

Table 5.2: P-value for 200000 points with frames of 1000 samples



(a) Autocorrelation for ANSD with LM

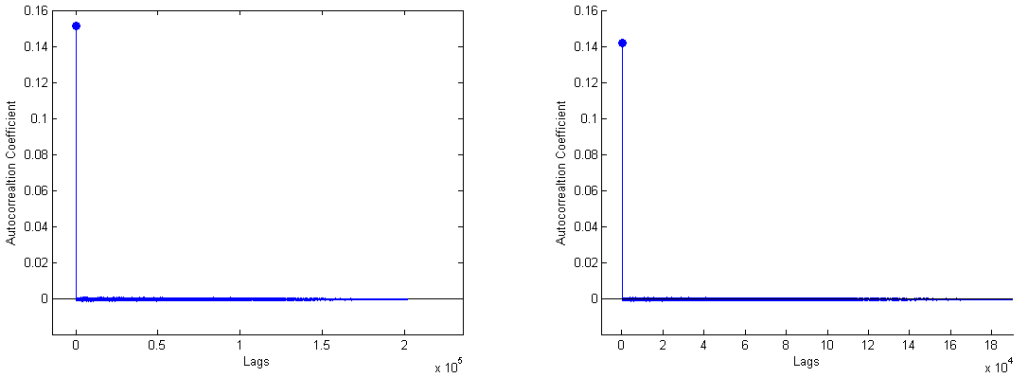


(b) Autocorrelation for ANSD with SPG

Figure 5.20: Autocorrelation of 200000 lags with frames of 2000 samples

	Ljung Box	Box Pierce	Frequency
Adaptive dither with LM	0.7825	1.0000	0.3905
Adaptive dither with SPG	0.7548	1.000	0.3689

Table 5.3: P-value for 200000 points with frames of 2000 samples



(a) Autocorrelation for ANSD with LM (b) Autocorrelation for ANSD with SPG

Figure 5.21: Autocorrelation of 200000 lags with frames of 4000 samples

	Ljung Box	Box Pierce	Frequency
Adaptive dither with LM	0.9692	1.0000	0.6224
Adaptive dither SPG	0.9985	1.0000	0.1091

Table 5.4: P-value for 200000 points with frames of 4000 samples

	Ljung Box	Box Pierce	Frequency
Adaptive dither with SPG	1.000	0.1446	0.2233

Table 5.5: P-value for 200000 points with frames of 10000 samples

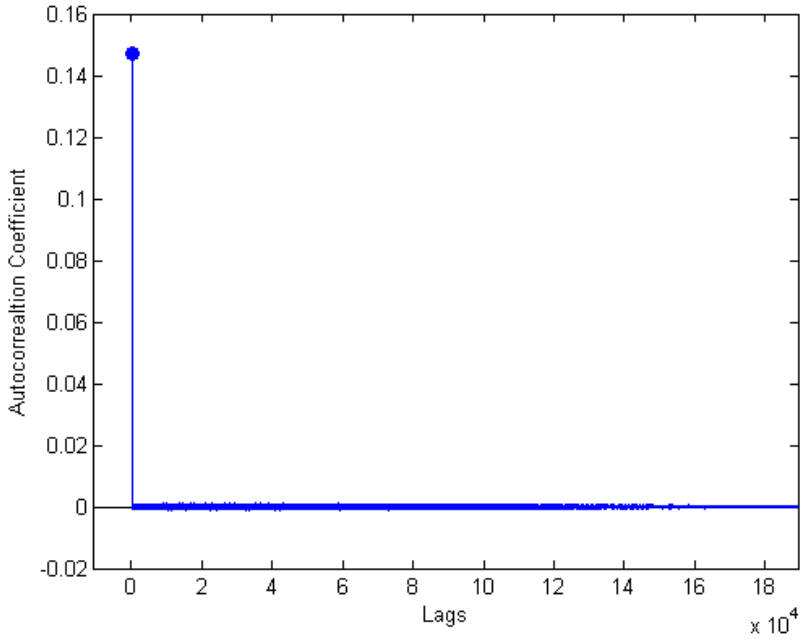


Figure 5.22: Autocorrelation of 200000 lags with frames of 10000 samples in SPG algorithm

5.3 Summary

The first section of this chapter has demonstrated that hypothesis testing can be used to determine if the total error of a re-quantized signal is white noise. Experiments in synthetic audio showed that Ljung-Box test is more accurate identifying non-white sequences than Box-Pierce and Frequency Test. In addition, the signal can be segmented, and each segment tested to determine the whiteness of the total error. Dither can be added to the segments that do not pass the tests. The second section presented experiments concerning to adaptive dithering using LM and SPG. In order to eliminate the correlation in LM, a circular autocorrelation estimator was used giving better results than linear autocorrelation estimator. The methods reaches the desired variance in the total error with an absolute error around 0, 01. Moreover, experiments show that the total error of adaptive dithering has

constant variance like the triangular PDF with higher SNR than classic dithering techniques

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

This research has demonstrated that hypothesis testing can be used effectively to determine if the total error of a re-quantized signal is white noise. The experiments indicate that for the 10% error level chosen, the tests are more sensitive than visual or audio inspection of the quantization error in determining if it is white noise or not. The application to real audio signal re-quantizing is straight forward. The signal can be segmented, and each segment tested for the whiteness of the quantization noise. Dither can be added to the segments that do not pass the tests. Of course, having one segment with no dither followed by another with dither may be undesirable, and also adding the typical full scale dither may not be needed. In addition to this, an application for segment dependent dithering was developed in C++ where a signal is re-quantized from 24 to 16 bits adding dither to the segments where the total quantization error is not white noise. The program has better execution time than Matlab and contains libraries to analize white noise conditions in time and frequency.

Adaptive dithering of one dimensional signal is a new technique which has been designed solving the non-linear system of equations of the autocorrelation of the total error. The derivatives of the system are approximated using a linear function. Iterative algorithms were then used to solve the resulting non-linear system. The methods were Levenberg Marquardt and Spectral Projected Gradient. For Spectral

Projected Gradient, each segment of the signal can be analyzed and the dither is obtained independently without showing a linear relationship between them. Also, the experiments show that the resulting variance in the total error using SPG reaches the desired variance with an absolute error of around 0.01. For the Levenberg Marquardt algorithm, the linear autocorrelation estimator is not useful for more than one frame. The reason for this is that the adaptive dither and the total error signals between frames are not independent causing a non-white noise signal, and consequently rejecting the null hypothesis of white noise. In order to eliminate the correlation in LM, a circular autocorrelation estimator was used giving better results than linear autocorrelation. With this estimator, the null hypothesis of the white noise test is not rejected. Moreover, different experiments changing the length of the frames and the number of bits in the re-quantized signal show that LM and SPG are accurate reaching the desired variance in the total error with a small error margin. Furthermore, the programs for SPG and LM were tested in C++ offering better performance than Matlab. Finally, the experiments show that $ANSD$ allows a total error signal with a constant variance from frame to frame and has a lower quantization noise variance than classic dithering techniques.

6.2 Future Work

In this research, adaptive dither is obtained in off-line mode, the method takes a lot of processing time for a small set of samples. The future work attempts to improved the processing time:

- Implementing an algorithm which allows finding adaptive dither in real time.
- Implementing the $ANSD$ in a parallel model which allows analyzing more data in less time. This implementation can have two schemes, the first consists of the implementation of the same program running on many platforms for different parts of the digital signal. And the second is to parallelize the optimization to improve the performance of the original program.

- To test adaptive dither and segment dependent dither for multidimensional problems.

APPENDICES

APPENDIX A
GRADIENTS AND JACOBIANS

A.1 Jacobian of Eq. 4.7

Let eq. 4.7 be defined as:

$$\mathrm{f}(\mathbf{d}) = \begin{pmatrix} f_0(\mathbf{d}) \\ f_1(\mathbf{d}) \\ f_2(\mathbf{d}) \\ f_k(\mathbf{d}) \\ f_{N-1}(\mathbf{d}) \end{pmatrix} = \begin{pmatrix} \epsilon[0]\epsilon[0] & + & \epsilon[1]\epsilon[1] & + & \dots & + & \epsilon[N-1]\epsilon[N-1] \\ 0 & + & \epsilon[0]\epsilon[1] & + & \dots & + & \epsilon[N-2]\epsilon[N-1] \\ 0 & + & 0 & + & \dots & + & \epsilon[N-3]\epsilon[N-1] \\ \vdots & & \ddots & & & & \\ 0 & + & \dots & + & 0 & + & \epsilon[0]\epsilon[N-1] \end{pmatrix} = \begin{pmatrix} N\sigma^2 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}$$

where ϵ depends on the dither signal. The first equation $f_0(\mathbf{d})$, is the biased variance, supposing that it has the partial derivative of $\epsilon[n]$ with respect to d_i, then the portion the gradient for all d_i of $f_0(\mathbf{d})$ is:

$$\nabla f_0(\mathbf{d})^T = \begin{bmatrix} 2\epsilon[0]\epsilon[0]\frac{d\epsilon[0]}{dd_0} \\ \vdots \\ 2\epsilon[N-1]\epsilon[N-1]\frac{d\epsilon[N-1]}{dd_{N-1}} \end{bmatrix}$$

For the equation $f_1(\mathbf{d})$ to the equation $f_{N-1}(\mathbf{d})$ it is happening that $\epsilon[i]$ in equation $f_n(\mathbf{d})$ is multiplied by $\epsilon[n+i]$ if and only if $0 < n+i < N-1$. In contrast, for the others $\epsilon[j]$ are multiplied by $\epsilon[j-n]$ and $\epsilon[j+n]$. So, the gradient for equation $f_1(\mathbf{d})$ to $f_{N-1}(\mathbf{d})$ is:

$$\nabla f_n(\mathbf{d})^T = \begin{bmatrix} \epsilon[0]\epsilon[n]\frac{d\epsilon[0]}{dd_0} \\ \vdots \\ \epsilon[n]\epsilon[n+i]\frac{d\epsilon[n]}{dd_n} + \epsilon[n]\epsilon[n-i]\frac{d\epsilon[n]}{dd_n} \\ \vdots \\ \epsilon[N-1-n]\epsilon[N-1]\frac{d\epsilon[N-1]}{dd_{N-1}} \end{bmatrix}$$

Finally, the jacobian is the matrix of gradients of the equations f_0 to f_{N-1}

$$J(f(\mathbf{d})) = \begin{bmatrix} \nabla f_0(\mathbf{d}) \\ \vdots \\ \nabla f_{N-1}(\mathbf{d}) \end{bmatrix}$$

A.2 Gradient of Eq. 4.9

Equation 4.9 is based in eq. 4.7 and is defined as:

$$F(\mathbf{d}) = \frac{1}{2}\sum_{i=0}^{N-1}(f_i(\mathbf{d}))^2$$

In order to obtain the gradient, partial derivatives with respect to d_i are taken as shown in the following equation:

$$\frac{dF(\mathbf{d})}{\partial d_i} = \sum_{i=0}^{N-1}\left(f_i(\mathbf{d})\left(\epsilon[n]\epsilon[n+i]\frac{d\epsilon[n]}{dd_n}I(n+i<N-1) + I(n-i>0)\epsilon[n]\epsilon[n-i]\frac{d\epsilon[n]}{dd_n}\right)\right),$$

where $I(a)$ is the indicator function. Finally, the gradient of F becomes:

$$\nabla F(\mathbf{d})^T = \begin{bmatrix} \frac{dF(\mathbf{d})}{\partial d_0} \\ \vdots \\ \frac{dF(\mathbf{d})}{\partial d_{N-1}} \end{bmatrix}$$

A.3 Jacobian of Eq. 4.14

Let eq. 4.14 be defined as :

$$\mathrm{f}(\mathbf{d}) = k\begin{pmatrix} \epsilon_{kn}[0]\epsilon_{kn}[0] & + & \epsilon_{kn}[1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-2]\epsilon_u[N-2] \\ \epsilon_u[N-1]\epsilon_{kn}[0] & + & \epsilon_{kn}[0]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-2]\epsilon_u[N-1] \\ \epsilon_u[N-2]\epsilon_{kn}[0] & + & \epsilon_u[N-1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_u[N-3]\epsilon_u[N-1] \\ \vdots & & \ddots & & & & \\ \epsilon_u[\frac{N}{2}]\epsilon_{kn}[0] & + & \epsilon_u[\frac{N}{2}+1]\epsilon_{kn}[1] & + & \dots & + & \epsilon_{kn}[\frac{N}{2}]\epsilon_u[N-1] \end{pmatrix} = k\begin{pmatrix} N\sigma^2 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}$$

It is necessary to calculate the gradients of the functions, then, the gradient of the equation $f_0(\mathbf{d})$ is defined as:

$$\nabla f_0(\mathbf{d})^T = \begin{bmatrix} 2\epsilon_u[0]\epsilon_u[0]\frac{d\epsilon_u[0]}{dd_0} \\ \vdots \\ 2\epsilon_u[N-1]\epsilon_u[N-1]\frac{d\epsilon_u[N-1]}{dd_{N-1}} \end{bmatrix}$$

For the equation $f_1(\mathbf{d})$ to the equation $f_{N-1}(\mathbf{d})$, the gradient is defined as:

$$\nabla f_n(\mathbf{d})^T = \begin{bmatrix} \epsilon_u[\frac{N}{2}]\epsilon_u[\frac{N}{2}+n]\frac{d\epsilon_u[\frac{N}{2}]}{dd_{\frac{N}{2}}} + \epsilon_u[\frac{N}{2}]\epsilon_u[\frac{N}{2}-n]\frac{d\epsilon_u[\frac{N}{2}]}{dd_0} \\ \vdots \\ (\ \epsilon_u[\frac{N}{2}+i]\epsilon_u[\frac{N}{2}+i+n]I(\frac{N}{2}+n+i<N-1)+ \\ \epsilon_u[\frac{N}{2}+i]\epsilon_u[\frac{N}{2}+i-n]I(i-n>0)+ \\ \epsilon_u[\frac{N}{2}+i]\epsilon_{kn}[\langle\frac{N}{2}+i+n\rangle_N]I(\frac{N}{2}+n+i>N-1)+ \\ \epsilon_u[\frac{N}{2}+i]\epsilon_{kw}[\frac{N}{2}+i-n]I(i-n<0)\ )\frac{d\epsilon[i]}{dd_{n1}} \\ \vdots \\ (\epsilon_u[N-1]\epsilon_u[\frac{N}{2}]+\epsilon_u[N-1]\epsilon_{kw}[\frac{N}{2}-1])\frac{d\epsilon[N-1]}{dd_{N-1}} \end{bmatrix}$$

Finally, the jacobian is the matrix of gradients of the equations f_0 to f_{N-1}

$$J(f(\mathbf{d})) = \begin{bmatrix} \nabla f_0(\mathbf{d}) \\ \vdots \\ \nabla f_{N-1}(\mathbf{d}) \end{bmatrix}$$

APPENDIX B
C CODE LIBRARY

The purpose of this section is to describe the set of main functions developed in this research. It is important to say that other functions like $DTFT$ are used directly of the GSL library.

- kstest: This function performs the goodness of fit Kolmogorov Smirnov test. This test measure the maximum difference between the sample cumulative distribution and the real cumulative distribution. In addition, this test is needed to perform the frequency test.
- qks: This function performs the q function in the Kolmogorov Smirnov test. In the general case $q = \sum_{k=1}^{\infty} e^{-2*k1^2*dst^2} * 2 * (-1)^{k1-1}$
- Freqtest: Perform the hypothesis test described in the section 4.1.2.
- Boxpierce: Perform the hypothesis test described in the section 4.1.1.1.
- Ljungbox: Perform the hypothesis test described in the section 4.1.1.2.

Furthermore, others functions were developed, but they have less relevance. Some of these functions are autocorrelation, variance and mean.

REFERENCE LIST

[1] L. G. Roberts. *Picture Coding Using Pseudo-Random Noise*. MIT, S.M. thesis, 1961.

[2] S. P. Lipshitz and J. Vanderkooy. Digital dither. *Proc. 81st Conv. Audio Eng. Soc., J. Audio Eng. Soc., vol. 34*, page 1030, Dec. 1986.

[3] B. Widrow. A study of rough amplitude quantization by means of nyquist sampling theory. *Circuit Theory, IRE Transactions on, Vol.3, Iss.4*, pages 266– 276, Dec 1956.

[4] Istvhn KollL B. Widrow and Ming-Chang Liu. Statistical theory of quantization. *IEEE Transaction on instrumentation and Measurement, VOL. 45, NO. 2*, pages 353–361, April 1996.

[5] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal, Vol. 27, pp. 379423, 623656*, October, 1948.

[6] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on communication technology*, 1964.

[7] Robert M. Gray and Thomas G. Stockham Jr. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

[8] Robert M. Gray. Quantization noise spectra. *IEEE Transactions on Information Theory*, 36(6):1220–1244, 1990.

[9] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory, VOL. 44, NO. 6*, pages 2325–2383, October 1998.

[10] John Vanderkooy R. A. Wannamaker, S. P. Lipshitz and J. Nelson Wright. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing, VOL. 48, NO. 2*, pages 499–516, February 2000.

[11] R. A. Wannamaker. *The Theory of Dithered Quantization,*. University of Waterloo, 1997.

[12] J. Vanderkooy and S. P. Lipshitz. Dither in digital audio. *J. Audio Eng. Soc. vol. 35*, page 966975, Dec. 1987.

[13] N. S. Jayant and L. R. Rabiner. The application of dither to the quantization of speech signals, bell syst. tech. j., vol. 51. *Maple Press*, JulyAug. 1972.

[14] J. Kim J. Kim and P. Jun. Dithered timing spread spectrum clock generation for reduction of electromagnetic radiated emission from high speed digital system. *Remote Sensing of Environment, Vol. 65*, pages 333–340, 1998.

[15] U Jonsson L. Ianelli, K. Johanson and F. Vasca. Analysis of dither in relay feedback systems. *Proceddings of the 41st IEEE Conference and Control*, December 2002.

[16] S. P. Lipshitz and R. Wannamaker. Quantization and dither: a theoretical survey. *J. Audio Eng. Soc. vol. 40*, pages 355–75, May 1992.

[17] C. F. Benitez-Quiroz and Shawn D. Hunt. Determining the need for dithering in one dimesional signals. *121 AES Convention*, 2006.

[18] G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelation in autoregressive-integrated moving average time series models. *J. of the American Statistical Association, Vol. 65, No. 332, pp. 1509-1526*, Dec., 1970.

[19] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika, 65, 2, pp. 297-30*, 1978.

[20] G. M. Jenkins and D. G. Watts. Spectral analysis and its applications. *Holden-Day*, 1968.

[21] C. F. Benitez-Quiroz and Shawn D. Hunt. Deconvolucion de gaussian mixture models, 2006.

[22] M. I. A. Lourakis. A brief description of the levenberg-marquardt algortihm implemented by levmar. *Lectures notes2,Foundation for research and technology*,

2004.

[23] H.B. Nielsen K. Madsen and O. Tingleff. Methods for non-linear least squares problems. *Lectures notes, Technical University of Denmark*, 2004.

[24] D. Rodriguez. *Review Notes on Advanced Signal Processing Algorithms*. University of Puerto Rico, 2005.

[25] Christian Kanzow, Nobuo Yamashita, and Masao Fukushima. Levenberg-marquardt methods for constrained nonlinear equations with strong local convergence properties.

[26] E. Birgin, J. Mart'inez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets, 1999.

[27] E. Birgin and J. Martinez. Large-scale active-set box-constrained optimization method with spectral projected gradients, 2002.

[28] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

[29] M.I.A. Lourakis. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. [web page] `http://www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004. [Accessed on 31 Jan. 2005.].

[30] L. D'Afonseca R. Biloti and S. Ventura. Open optimization library. [web page] `http://ool.sourceforge.net//`, June 2005.