ADAPTIVE DITHERING OF ONE DIMENSIONAL SIGNALS

Carlos Fabian Benitez-Quiroz and Shawn Hunt

Univeristy of Puerto Rico, Mayaguez Campus

ABSTRACT

This work presents a novel dithering algorithm for one dimensional signals. Standard dithering with a triangular probability density white noise increases the noise level by 6dB. This technique is numerically based instead of statistically based, and the increase in noise level can be specified as low as 3dB. The problem is reduced to a set of non-linear simultaneous equations that are solved using any number of optimization techniques. In this paper, a solution using Levenberg-Marquardt is shown and is compared to standard dithering with both real and synthetic signals.

Index Terms— Dither techniques, Signal quantization, Adaptive signal processing

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. 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). This assumption does not hold for all cases, particularly when the signal amplitude is small compared to the quantization step. In this case rounding or truncating a signal can introduce various undesirable artifacts, namely, additional harmonics related to the signal being re-quantized.

To avoid these unwanted harmonics, dither is generally added to the signal being quantized. The classic dithering of one dimensional signals consists of a white noise signal whose purpose is to ensure that the quantization error is uncorrelated with the signal being quantized. Also, it is assumed that the dither signal is independent of the input. In non-subtractive dithering this independence assumption is never met because the quantization error signal is dependent on the signal being quantized. However, some techniques can make the first and second statistical moments independent of the input. For instance, triangular PDF dither ($TPDF$) has this property. It has a higher noise level than some other methods, but psycho-acoustical tests show that the constant noise level produced by having the second moment independent is preferred by users. On the other hand, the main disadvantage of adding dither is that since it is a noise signal, the signal to noise ratio (SNR) of the re-quantized signal is lowered. The work presented here has been to develop an adaptive dither with the same properties as $TPDF$ with higher SNR than the classic dithering method.

This paper is organized as follows. In section 2 the theoretical foundation of quantization and dithering is presented. Section 3 presents the new technique for finding signal dependent dither using constrained optimizations algorithms. Section 4 presents experiments comparing this method with classic dithering using real and synthetic data. The conclusions and future work are presented in the last section.

2. RE-QUANTIZATION AND DITHERING

2.1. Re-quantization

This work is focused on dithering for discrete signals when reducing the number of bits. This process of lowering the number of bits is referred to as re-quantization. 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, and ε the quantization noise. The simplest method of lowering the resolution is rounding where the number is approximated by the nearest integer. Since the numbers are in 2s complement,

This work was partially supported by the Engineering Research Center Program of the National Science Foundation under Award Number EEC-9986821

the rounding operation is

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

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

In classic quantization model, the quantization noise is assumed white, so its autocorrelation is an impulse at lag zero. The Fourier transform of the autocorrelation is the Power Spectral Density, and so has constant amplitude.

2.2. Dithering

The purpose of the dither signal is to ensure that the quantization error is uncorrelated with the signal being quantized.

Non-subtractive dither (NSD) is commonly used for one dimensional signals. In this application dither is added, and the signal is then re-quantized. In NSD, the input signal can be modeled as a linear mixture model given by:

$$x[n] = Q(x[n] + d[n]) + \varepsilon[n] = x_q[n] + \varepsilon[n], \quad (1)$$

where ε is the total error. Lipshitz et. al. [1] have shown that the total error PDF $p_\varepsilon(\varepsilon)$ is always dependent of input as can be seen in the following equation:

$$p_\varepsilon(\varepsilon) = \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x) p_x(x) dx, \quad (2)$$

where $p_x(x)$ is the input PDF and $p_{\varepsilon|x}(\varepsilon, x)$ is defined by:

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k-\infty}^{\infty} \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_d(w-x) dw. \quad (3)$$

Taking the Fourier transform of Eq. 2 and using the convolution theorem, the CF of ε is given by:

$$C_\varepsilon(f) = \sum_{k=-\infty}^{\infty} sinc\left(f - \frac{k}{\Delta}\right) C_d\left(f - \frac{k}{\Delta}\right) C_x\left(-\frac{k}{\Delta}\right), \quad (4)$$

where C_ε, C_x and C_d denotes the characteristic function of the total error, the input and the dither signal respectively. Based on this formula, Lipshitz in [2] and [3] demonstrates that if

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

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 respect to 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 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 for many one dimensional signal applications.

3. ADAPTIVE DITHERING

Let the quantized signal be defined as:

$$x_q[n] = Q\left(\frac{x[n]}{2^m} + d[n]\right) 2^m, \quad (5)$$

where 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 let $\frac{x[n]}{2^m} = x_i[n] + x_f[n]$ where $x_i[n]$ represents the integer part (possibly zero) and $x_f[n]$ the fractional part.

If the input to Q has integer parts, they are not affected by the quantizer. This is used to re-write Eq. 5 as Eq. 6. $x_f[n]$ is fractional and must be evaluated with the dither $d[n]$ which is unknown so Eq. 5 becomes:

$$x_q[n] = (x_i[n] + Q(x_f[n] + d[n]))2^m. \quad (6)$$

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

$$\epsilon[n] = \frac{x[n] - x_q[n]}{2^m}. \quad (7)$$

Replacing Eq. 6 by Eq. 7, the following expression is obtained:

$$\epsilon[n] = x_f[n] - Q(x_f[n] + d[n]). \quad (8)$$

At this point, it is clear that the total error depends only on the fractional part of $\frac{x[n]}{2^m}$. Hence, the integer part can be ignored in the following sections.

The dither is obtained using the autocorrelation of the total error. The signal is segmented into frames of length N and the total error is iteratively calculated finding the dither $d[n]$ which minimizes the difference between the actual and desired autocorrelation. The solution is improved using the circular autocorrelation estimator [4] instead of the linear autocorrelation estimator. The circular autocorrelation estimator is used to measure the linear and circular relationship between two samples. Let the circular autocorrelation estimator at lag L be defined as:

$$f_L(\mathbf{d}) = \sum_{n=0}^{N-1} \varepsilon(\langle n - L \rangle_N)\varepsilon(n), \tag{9}$$

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. This means the system is underdetermined (i.e, less equations than variables) and it is necessary to add additional equations in order to solve it as an overdetermined problem. Assuming that the first frame of length N hs been solved and the total error computed, the next frame is solved 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. 10 presents the system of equations for the LM algorithm.

$$\mathrm{f}(\mathbf{d}) = \begin{bmatrix} \epsilon_{kn}[0]\epsilon_{kn}[0] + & \epsilon_{kn}[1]\epsilon_{kn}[1] + & \ldots + \epsilon_u[N-2]\epsilon_u[N-2] \\ \epsilon_u[N-1]\epsilon_{kn}[0] + & \epsilon_{kn}[0]\epsilon_{kn}[1] + & \ldots + \epsilon_u[N-2]\epsilon_u[N-1] \\ \epsilon_u[N-2]\epsilon_{kn}[0] + & \epsilon_u[N-1]\epsilon_{kn}[1] + & \ldots + \epsilon_u[N-3]\epsilon_u[N-1] \\ \vdots & & \ddots \\ \epsilon_u[\frac{N}{2}]\epsilon_{kn}[0] + & \epsilon_u[\frac{N}{2}+1]\epsilon_{kn}[1] + & \ldots + \epsilon_{kn}[\frac{N}{2}]\epsilon_u[N-1] \end{bmatrix} = \begin{bmatrix} N\sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{10}$$

where ϵ_u and ϵ_{kn} are the unknown and known error samples respectively. Furthermore, the system is clearly non-linear because $\epsilon[n]$ depends on $Q(x)$ which is non-linear. The system in Eq. 10 can be solved using any number of optimization algorithms. To linearize the system, popular optimization algorithms use derivatives or a finite approximation of them. In this case, $Q(x)$ is not differentiable, so a smooth continuous function is needed for the linearization. The simplest estimator of $Q(x)$ is to linearize $\hat{Q}(x) = x$, so $\frac{d\hat{Q}(x)}{dx} = 1$. This approximation has given good results as can be seen in section 4. The following section presents an algorithm based on Levenberg Marquardt that solves the optimization problem in a constrained space.

3.1. Box Constrained Levenberg-Marquardt Algorithm

This algorithm was proposed by Kanzow et. al. in [5]. The algorithm is called a projected Levenberg-Marquardt (PLM) method because $\mathbf{d}$ is a projection onto the feasible 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 into the desired region. The update sequence is given by:

$$\mathbf{d}^{k+1} = P_X\left(\mathbf{d}^k + \boldsymbol{\alpha}\right), \tag{11}$$

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 [5] to have a decreasing error.

4. EXPERIMENTS AND RESULTS

4.1. Experiments using Synthetic Signals

The following experiment uses 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 PLM and the desired variance. The total error variance of the adaptive dither technique described in section 3 is compared with the variance of classical dithering 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 PLM 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 4000 samples.

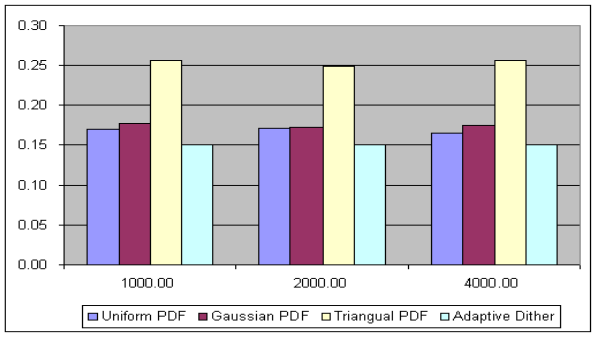

Fig. 1. Total error variance for the re-quantized signal for different frame lengths

Figure 1 shows that the absolute error between the desired and sample variance for the adaptive dither is around 0.01. Both $GPDF$ and $RPDF$ dithers can have variances that change depending of the input signal in contrast with $TPDF$ dither in which the variance is independent of the signal. In the case of the adaptive dithering, the desired variance was very close to that specified.

The previous experiment measures the variance of the total error when re-quantizing to 16 bits. The next experiment changes the number of bits to see if adaptive dither is effective when requantizing to various bit depths. The signal is again 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 2 presents the variance of the total error at different quantization levels. The results show that adaptive dither reaches the specified variance. Furthermore, Figure 2 shows that for $RPDF$ and $GPDF$ the total error variance changes depending of the quantization level.

In solving the nonlinear equations for the adaptive dither,

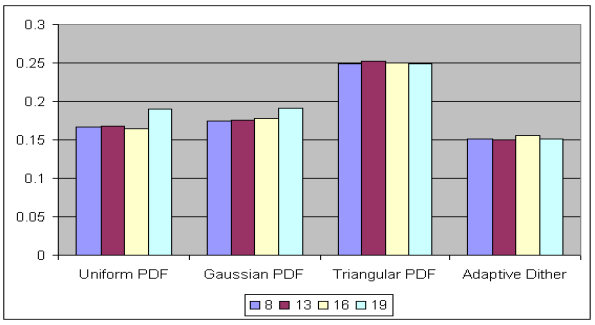

Fig. 2. Variance of the total error at different quantization levels

the desired variance must be specified. In the next experiment, this level is changed to see how well the algorithm performs for variance levels between 0.11 and 0.3.

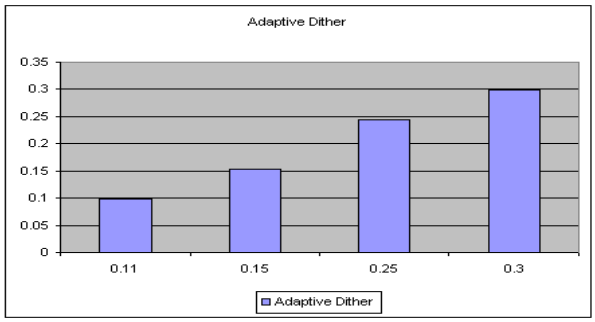

Fig. 3. Actual vs target variance of the total error for different target levels

Figure 3 shows that the sample variance of the total error reaches the desired variance.

4.2. Experiments using Real Signals

The purpose of the following experiment is to test adaptive dithering using real one dimensional signals when using different segment lengths. The experiments used real audio signals with 24 bit precision and a 44.1 kHz sampling rate. This experiment uses a signal with 200,000 samples segmented into frames of length 1000, 2000, and 4000 samples. The desired variance in the total error is set at 0.150. The original audio file with 24 bit precision, and is re-quantized to 16 bits. Figure 4 shows the total error variance for the different segment lengths. As seen in this figure, the desired variance is reached. The resulting noise was analyzed as in [6] to determine if the total error is white noise. It was classified as white noise by all the methods in that paper.

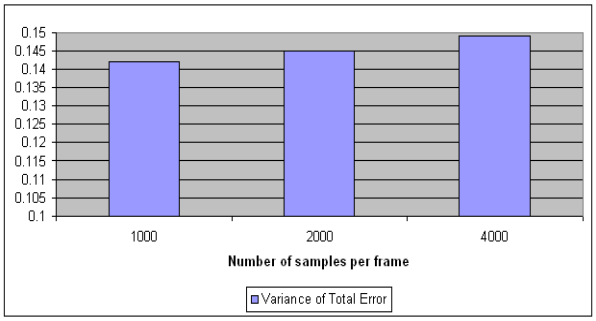

Fig. 4. Variance of the total error for different frame lengths

5. CONCLUSIONS

Adaptive dithering of one dimensional signal is a new technique which has been designed by solving a system of nonlinear equations resulting from the autocorrelation of the total error. The derivatives of the system are approximated using a linear function, and a Projected Levenberg-Marquards was then used to solve the resulting non-linear system. In different experiments changing the length of the frames and the number of bits in the re-quantized signal the adaptive dithering algorithm reaches the specified variance for the total error within a small error margin. Finally, the experiments show that adaptive dithering allows a total error signal with a constant variance from frame to frame and has a lower total error variance than classic dithering techniques.

6. REFERENCES

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

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

[3] 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*, pp. 499–516, February 2000.

[4] D. Rodriguez, *Computational Signal Processing and Sensor Array Signal Algebra: A Representation Development Approach*, University of Puerto Rico, 2002.

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

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