COEFFICIENT DITHER IN FIXED-POINT FIR DIGITAL FILTERS

Sourav R. Dey and Alan V. Oppenheim

Massachusetts Institute of Technology
Digital Signal Processing Group
77 Massachusetts Avenue, Cambridge MA 02139
sdey@mit.edu, avo@mit.edu

ABSTRACT

This paper presents dynamic coefficient dither as a method to mitigate coefficient quantization error in FIR filters. The uncorrelated, shapeable noise from dithered coefficients may be preferable, in certain contexts, to the frequency response distortion due to static fixed-point implementation. A theoretical framework for the analysis of coefficient dither is developed. Performance, fundamental tradeoffs, and optimization of dithered filter implementations are discussed. Oversampling is shown to improve the SNR. The theoretical results are validated with numerical experiments.

Index Terms— dither, fixed-point filters, coefficient quantization, vector binary processes, noise-shaping

1. INTRODUCTION

Dither has long been used as an alternative to static nonlinearities in signal quantization, but its use in mitigating coefficient quantization error in fixed-point digital filters is much less developed. Dithered filters have been proposed before in [1] based on an empirical treatment. In this paper, we present a theoretically based development of dithered Direct Form FIR filters.

In our development, we assume that a desired FIR impulse response, $\{b_i\}_{i=0}^{N-1}$, with N continuous-valued coefficients, is given. Fixed-point implementation leads to distortion from this desired response [2]. A number of techniques have been developed to minimize this distortion, from optimal bit-allocation programs [3, 4] to special structures, including cascaded-integrator-comb (CIC) filters [5]. Each has shortcomings such as design difficulties and limited re-configurability. Moreover, static fixed-point implementations typically lead to distortion of the frequency response. In certain applications, especially perceptual ones, this distortion can be unacceptable.

In this paper, we present an approach inspired by [6] based on dynamically dithering fixed-point coefficients that mitigates the frequency distortion caused by static coefficient quantization. Coupled with over-sampling and properly designed dither correlation, we show that coefficient dithering can use coarse fixed-point coefficient representations to achieve the performance of filters with much more finely quantized coefficients. In the limit, coefficient dithering can be used to implement high-quality filters with one-bit, multiplier-less coefficients. Such multiplier-less filters could be useful in a number of applications, including ones that require low latency or small VLSI chip area.

Section 2 introduces the standard dithered filter model and the four types of dithered filters. Section 3 develops Type I standard dithered filters in detail. Section 4 introduces the oversampled dither filter model. Section 5 develops Type I oversampled dithered filters in detail. Section 6 includes experiments from numerical simulations to validate the theoretical findings. Coefficient dithering is shown to offer a powerful alternative to static fixed-point coefficient quantization.

2. STANDARD DITHERED FILTER MODELS

In this paper, we assume uniform coefficient quantization with a fixed step-size Δ. The analysis can be extended to non-uniform coefficient quantization in a straightforward manner. Figure 1 illustrates a standard Direct Form FIR dithered filter. Each coefficient in the tapped delay-line is a random process, $b_i[n]$, of the form:

$$b_i[n] = Q(b_i) + s_i \Delta d_i[n] \tag{1}$$

where $Q(b_i)$ is the quantized coefficient, $s_i = \text{sgn}(b_i - Q(b_i))$, and $d_i[n]$ is a binary dither process which takes the values 0 or 1 at each n. The mean of $d_i[n]$ is constrained such that the dither tap takes the desired continuous-value on average, i.e. $E\{b_i[n]\} = b_i$. Mathematically, this implies the constraint:

$$E\{d_i[n]\} = \mu_i = \frac{(b_i - Q(b_i))s_i}{\Delta} \tag{2}$$

With the constraint of Eqn.(2), the dithered coefficient can be expressed as:

$$b_i[n] = b_i + s_i \Delta \tilde{d}_i[n] \tag{3}$$

where $\tilde{d}_i[n] = \{-\mu_i, 1 - \mu_i\}$ is a zero mean process such that $d_i[n] = \mu_i + \tilde{d}_i[n]$. Using the decomposition of Eqn.(3), the output of the dithered filter, $\hat{y}[n]$, can be expressed as:

$$\hat{y}[n] = \underbrace{\sum_{i=1}^{N-1} b_i x[n-i]}_{y[n]} + \underbrace{\Delta \sum_{i=0}^{N-1} s_i \tilde{d}_i[n] x[n-i]}_{e[n]} \tag{4}$$

The first term, $y[n]$, is the desired output from a Direct Form FIR filter with continuous-valued taps b_i. The second term, $e[n]$, is dither noise. In our analysis, we assume that the input and dither are independent wide sense stationary (WSS) random processes. From the constraint of Eqn.(2), the dithered filter has the desired response on average, i.e. $E\{e[n]\} = 0$. Because the dither and input are independent, the dither noise is statistically uncorrelated with the desired output. It can be shown to be a wide sense stationary (WSS) random

This work was supported in part by the NDSEG Fellowship, TI Leadership University, BAE Systems, and Lincoln Labs.

978-1-4244-2941-7/08/$25.00 ©2008 IEEE

process with auto-correlation:

$$R_{ee}[m] = \Delta^2 \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E\left\{\tilde{d}_i[n]\tilde{d}_j[n+m]\right\} s_i s_j E\left\{x[n-i]x[n+m-j]\right\} \quad (5)$$

This auto-correlation can be shaped and reduced by properly designing the dither correlation. This is the benefit of dithering. The zero-mean, uncorrelated, shapeable dither noise may be less disturbing, especially in perceptual applications, than the frequency response distortion that could result from static fixed-point implementations.

We use the SNR, defined below, as the error metric in this paper.

$$\text{SNR} = \frac{E\{y^2[n]\}}{E\{e^2[n]\}} \quad (6)$$

The coefficient dither can be viewed as a vector binary process:

$$\mathbf{d}[n] = \begin{bmatrix} d_0[n] & d_1[n] & \cdots & d_{N-1}[n] \end{bmatrix}^T \quad (7)$$

There are four forms of coefficient dither depending on the two-dimensional correlation properties of $\mathbf{d}[n]$, in time and across the taps. These are summarized in Table 1. In the simplest form, referred to as Type I, the dither across taps is independent and a memoryless Bernoulli process for each tap. Type I dither is developed in detail in Section 3.

Correlation across the taps can be used to improve performance. In Type II dither, each tap is a Bernoulli process but the correlation between the taps can be designed to increase the SNR. Intuitively, the coefficient dither can be designed so that there is partial error cancellation at the accumulator. Design is difficult because of the binary nature of $\mathbf{d}[n]$. Type II dither is not developed in detail in this paper. The interested reader is referred to [7] for a detailed development.

Coefficient dither can also be correlated in time. Time correlation can be used to frequency-shape the error spectrum in a desirable manner. This may be useful in perceptual applications such as audio, since the ear is less sensitive to high frequency noise. It is particularly useful in oversampled coefficient dithering, introduced in Section 4. Similar to Type II dither, the design of time correlation is difficult because of the binary nature of $\mathbf{d}[n]$. Neither Type III and Type IV dither are developed in detail in this paper. Preliminary development of these forms of dither can be found in [7].

	Time-Independent	**Time-Correlated**
Tap-Independent	Type I	Type III
Tap-Correlated	Type II	Type IV

Table 1. Forms of coefficient dither depending on correlation.

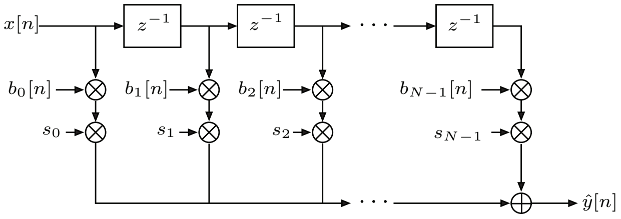


Fig. 1. Standard coefficient dithering model.

3. STANDARD TYPE I DITHERED FILTERS

Standard Type I dithered filters are the simplest form of dithered filters with $\mathbf{d}[n]$ independent in time and across the taps, i.e. $E\{d_i[n]d_j[n+m]\} = \sigma_i^2\delta_{ij}\delta[m]$. Substituting into Eqn.(5), the dither noise can be shown to be white with auto-correlation:

$$R_{ee}[m] = \Delta^2 \sum_{i=0}^{N-1} \sigma_i^2 R_{xx}[0]\delta[m] = \mathcal{E}_1\delta[m] \quad (8)$$

where $\mathcal{E}_1 = E\{e^2[n]\}$ is the MSE. Recall that in a binary process the mean fixes the variance as $\sigma_i^2 = \mu_i(1-\mu_i)$. Substituting the mean constraint of Eqn.(2) into the expression for the variance into Eqn.(8), the MSE can be expressed as:

$$\mathcal{E}_1 = \Delta R_{xx}[0]\sum_{i=0}^{N-1}|b_i - Q(b_i)| - R_{xx}[0]\sum_{i=0}^{N-1}|b_i - Q(b_i)|^2 \quad (9)$$

The b_i are fixed by the desired continuous-valued filter and $Q(b_i)$ and Δ are fixed by the coefficient quantization levels. Consequently, in its simplest formulation, there are no designable parameters for Type I dithered filters and the MSE is fixed by the desired filter and coefficient quantization.

In a more advanced formulation, we can achieve a degree of freedom by scaling the desired continuous-valued taps to b_i/K and adding a continuous-valued scaling K after the filter. In this case, the filter still has the desired response on average, but the MSE can be reduced by choosing the scaling appropriately. Naively, it may seem that the MSE should be independent of scaling K, but this is not the case. For example, assume that the desired continuous-valued filter has three taps, $\mathbf{b} = \begin{bmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix}$ and that the quantization levels are the integers, $Q(b_i) = \{\ldots,-1,0,1,\ldots\}$. There are an unlimited number of $(\frac{\mathbf{b}}{K}, K)$ pairs that can be used to implement this filter. Two possible pairs are:

$$(\frac{\mathbf{b}_1}{K_1}K_1) = (\begin{bmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix}, 1);\ (\frac{\mathbf{b}_2}{K_2}, K_2) = (\begin{bmatrix}1 & 1 & 1\end{bmatrix}, \frac{1}{2}) \quad (10)$$

The first pair has dithered taps that switch between 0 and 1 and has a non-zero MSE. In contrast, the second pair can be implemented perfectly as a static filter so the MSE is zero. Clearly the second pair is a better dithered implementation and the choice of scaling matters.

Formally, the scaling design can be expressed as a constrained optimization:

$$\begin{array}{ll} \underset{K}{\text{minimize}} & K^2\Delta R_{xx}[0]\sum_{i=0}^{N-1}|\frac{b_i}{K} - Q(\frac{b_i}{K})| \\ & -K^2 R_{xx}[0]\sum_{i=0}^{N-1}|\frac{b_i}{K} - Q(\frac{b_i}{K})|^2 \\ \text{subject to} & \frac{\max\{|b_i|\}}{K} \leq \max\{|Q(b_i)|\} \end{array} \quad (11)$$

The constraint ensures that the scaling does not saturate the coefficient quantization levels. This optimization can be solved numerically for the optimal scaling K. For the special case of one-bit coefficient quantization, where $Q(\frac{b_i}{K}) = \{-1, 0, 1\}$, the design problem can be expressed as:

$$\begin{aligned} \underset{K}{\text{minimize}} \quad & KR_{xx}[0]\sum_{i=0}^{N-1}|b_i| - R_{xx}[0]\sum_{i=0}^{N-1}|b_i|^2 \\ \text{subject to} \quad & K \geq \max\{b_i\} \end{aligned} \tag{12}$$

The optimal scaling in this special case is to choose K as the minimal feasible value, $K^* = \max\{|b_i|\}$. Intuitively, each random tap is a source of noise, so the optimal solution is to make the tap with maximal absolute value statically equal to 1 or -1. For a more complete study of scaling optimization for Type I filters, the reader is referred to [7].

Standard dithered filters do not have a better SNR than an optimal static implementation. Intuitively, since there are only a finite number of static fixed-point implementations, these exists a static implementation that has the minimum MSE. Switching randomly, using dither, to other non-minimum configurations can only increase the MSE. The benefit of dither comes from the fact that the form of the error is potentially less disturbing not a lower MSE.

Type I MSE is dependent on the desired continuous-valued filter we are trying to implement. Certain filters, like one where the coefficients are all identical, e.g. $\mathbf{b} = \begin{bmatrix} b & b & b \end{bmatrix}$, can be perfectly implemented using a a fixed-point implementation and has zero MSE. Other filters have a higher MSE. For the special case of one-bit coefficient quantization, the worst-case MSE for a given length N can be shown to be:

$$\mathcal{E}_1(N) \leq \frac{R_{xx}[0]}{2}\left(\sqrt{N-2+\frac{1}{N}} - 1 + \frac{1}{\sqrt{N}}\right) \tag{13}$$

The derivation is omitted for the sake of brevity. The interested reader is referred to [7] for a complete derivation of this result.

As implied by Eqn.(13), Type I MSE increases as the number of taps N increases. Intuitively this is because with more taps there are more noise-sources injecting error into the filter. The growth of the error depends on the ideal filter specification. However, with more taps we can implement a better filter on average, i.e. we are closer to the ideal filter specification. There is no ideal operating length, rather the system designer must choose a suitable operating length by trading off between the dither noise and filter approximation error. Section 6 illustrates this tradeoff for a specific example.

4. OVERSAMPLED DITHERED FILTER MODELS

In addition to dithering in the standard Direct Form structure, we develop an ovresampled Direct Form structure illustrated in Figure 2. There are three major differences from the standard structure. First, the tapped-delay line is preceded by an upsampling stage. As illustrated, The input, $x[n]$, is expanded by a factor of L and interpolated with, $G_u(e^{j\omega})$, a LPF with gain L and cutoff π/L.

Secondly, the tapped-delay line is expanded, i.e. the unit delays are replaced with L-element delays. The dithered taps are the same as in the standard Direct Form structure thought, i.e. the assumptions of Eqns.(1) through (3) still hold. Note that the tapped delay-line has N non-zero tap processes, not LN. This is important because, as discussed in Section 3, the MSE scales with the number of non-zero tap processes. By fixing the number of taps to N, the MSE is fixed independent of the rate L.

Thirdly, in the oversampled structure, the tapped delay-line is followed by a down-sampling stage. As illustrated in Figure 2, the output of the tapped delay line is anti-aliased with a unity-gain LPF filter, $G_d(e^{j\omega})$, with cutoff π/L, and then compressed by L.

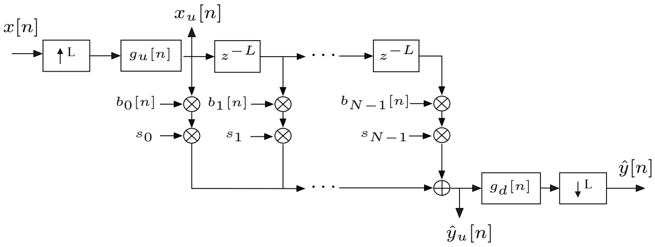


Fig. 2. Oversampled coefficient dithering model.

Using the decomposition of Eqn.(3) , the output before downsampling, $\hat{y}_u[n]$, can be expressed as:

$$\hat{y}_u[n] = \underbrace{\sum_{i=1}^{N-1} b_i x_u[n-iL]}_{y_u[n]} + \underbrace{\Delta\sum_{i=0}^{N-1} s_i\tilde{d}_i[n]x[n-iL]}_{e_u[n]} \tag{14}$$

After downsampling, the first term become $y[n]$, the desired output from a Direct Form filter with continuous-valued taps b_i. Similarly, the upsampled dither noise, $e_u[n]$, is downsampled into the dither noise $e_L[n]$. Mathematically, the output $\hat{y}[n]$ can be expressed as:

$$\hat{y}[n] = y[n] + e_L[n] \tag{15}$$

As in standard dithered filters, the output dither noise can be shown to be zero-mean, WSS process that is uncorrelated with the desired output. Its power spectrum can be expressed as:

$$S_{e_L e_L}(e^{j\omega}) = \frac{1}{L}S_{e_u e_u}(e^{j\omega/L})|G_d(e^{j\omega/L})|^2 \tag{16}$$

where $S_{e_u e_u}(e^{j\omega})$ is the power spectrum of $e_u[n]$ from Eqn.(14). From Eqn.(16), we see that the MSE is determined by the power of $e_u[n]$ in the passband of $G_d(e^{j\omega})$, i.e. $|\omega| < \pi/L$. The autocorrelation of $e_u[n]$ can be expressed as:

$$R_{e_u e_u}[m] = \Delta^2 \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E\{\tilde{d}_i[n]\tilde{d}_j[n+m]\} \\ s_i s_j R_{x_u x_u}[m + L(i-j)] \tag{17}$$

where $x_u[n]$ is the upsampled input. Similar to standard dithered filters, there are four types of oversampled dithered filters depending on dither correlation. In this paper we only discuss Type I in detail. Oversampling is shown to provide an L-fold SNR gain over standard Type I dithered filters. Frequency-shaped dither, both Type III and Type IV, is particularly useful for oversampled dithered filters because the dither noise can be shaped into the stop-band of $g_d[n]$. This can significantly increase the SNR at the output using an effect similar to sigma-delta noise-shaping. Though frequency-shaped dither is not discussed in this paper, the interested reader is referred to [7] for a preliminary development.

5. OVERSAMPLED TYPE I DITHERED FILTERS

In Type I oversampled dithered filters $\mathbf{d}[n]$ is independent in time and across taps, i.e. $E\{\tilde{d}_i[n]\tilde{d}_j[n+m]\} = \sigma_i^2\delta_{ij}\delta[m]$. Substituting into Eqn.(17), the upsampled dither noise auto-correlation can be expressed as:

$$R_{e_u e_u}[m] = \Delta^2 \sum_{i=0}^{N-1} \sigma_i^2 s_i^2 R_{xx}[0]\delta[m] \tag{18}$$

Which is white and equivalent to the standard Type I dither noise auto-correlation, $R_{ee}[m]$, from Eqn.(8). This is intuitively sensible because the oversampled dither noise is from the same number of noise sources with the same means.

Since the noise spectrum is identical, the optimal design of Type I oversampled dithered filters is the same as that for standard ones. Specifically, if allowed, the optimal scaling, K, is the same. Oversampling does not add any new degrees of freedom that can be exploited using Type I dither.

At the output, the Type I dither noise spectrum is:

$$S_{e_L e_L}(e^{j\omega}) = \frac{\mathcal{E}_1}{L} \tag{19}$$

Accordingly, the SNR is L times higher than that of a standard implementation:

$$\text{SNR}_L = L\frac{E\{y^2[n]\}}{\mathcal{E}_1} = L \cdot \text{SNR}_1 \tag{20}$$

Intuitively, oversampled coefficient dithering can be interpreted as the average of L standard dithered filters running independently. The averaging reduces the dither noise variance while keeping the desired output unchanged.

Note that there is no theoretical limit to oversampling gain. It can be used to arbitrarily improve the output SNR – well beyond that of a static fixed-point implementation. In addition, the dither noise still remains white and uncorrelated with the input. In practice though, the hardware constraints will impose a limit, i.e. with larger L the tapped delay-line must be longer and dither processes must run faster. All in all, oversampled dithered filters offer a powerful alternative to static fixed-point coefficient quantization.

6. NUMERICAL EXPERIMENTS

In this section, we illustrate the performance of Type I dithered filters using an example. In our example, the input $x[n]$ is a WSS DT ARMA process generated by shaping white Gaussian noise through a filter:

$$G(z) = \frac{(z-z_0)(z-z_0^*)}{(z-p_0)(z-p_1)} \tag{21}$$

with $z_0 = e^{j\pi/2}$, $p_0 = 0.9$, and $p_1 = -0.6$. The desired continuous-valued filter is a $N = 33$ tap, linear-phase, Parks-McClellan low-pass filter designed using the specifications: $H(e^{j\omega}) = 1$ for $\omega_p = [0, 3\pi/16]$ and $H(e^{j\omega}) = 0$ for $\omega_s = [5\pi/16, \pi]$.

Our goal is to implement a fixed-point implementation of this filter where each coefficient has been quantized to one-bit. Mathematically, in our dithered implementation this implies that taps can only take binary values $b_i[n] = \{0, 1\}$ and $\Delta = 1$. Such a filter is essentially multiplier-less, the coefficient multiplies are replaced with switches.

We design a scale optimized Type I dithered implementation by solving the optimization of Eqn.(12) . We implement a standard Type I dithered filter and an oversampled Type I dithered filter with $L = 4$. For each, we simulate the dithered filter in MALTAB and generate two million samples of the dither noise $e[n]$ or $e_L[n]$. Periodogram averaging with a Hamming window of size 2048 with 50% overlap is used to approximate 2048 samples of the power spectrum $S_{ee}(\omega)$. The MSE is estimated numerically by averaging the squared difference between $y[n]$, the desired output of the continuous-valued filter, and $\hat{y}[n]$ output of the dithered filter.

Figure 6 illustrates the results of Type I standard dithered implementation. Figure 6(a) illustrates $S_{ee}(e^{j\omega})$. As expected, it is white with height given by Eqn.(8). Figure 6(b) illustrates a section of the output, $\hat{y}[n]$, in the time domain along with the desired output, $y[n]$. The output of the dithered filter is a degraded version of the desired output $y[n]$. For this example the SNR is 7.18 dB.

Figure 7 illustrates the same results for the Type I oversampled dithered implementation. As expected, the error spectrum is still white, but the noise floor has been reduced due to oversampling gain. There is a small amount of distortion in the error spectrum near $\omega = \pi$ due to the use of non-ideal rate-conversion filters. In the time-domain, the output of the oversampled BRF, $\hat{y}[n]$, more closely follows the desired output, $y[n]$. The SNR is 13.52 dB, illustrating a 6.34 dB oversampling gain over the standard Type I implementation.

Figure 4(a) illustrates the Type I standard dithered filter SNR scaling for this particular example as a function of N. As expected, the MSE grows on the order $O(\sqrt{N})$. The error scaling is slower than the worst-case, Eqn.(13) by a large multiplicative factor.

As noted in Section 3, even though the MSE increases with N, with more taps we can implement a better filter. For this Parks-McClellan example, we can measure the the filter performance using the max ripple error. Figure 5 illustrates the max ripple error as a function of N for our filter specifications. It decays quickly with N. The system designer must make a tradeoff between the max ripple error in Fig.5 and the SNR in Fig.4 to choose an operating point.

Figure 3 illustrates the SNR as function of oversampling rate, L, for this example. As expected, the SNR grows as $10\log_{10} L$ on the dB plot. Figure 3 also illustrates the SNR of a static fixed-point filter with one-bit coefficients. It is constant as a function of L. With about 4x oversampling we can outperform the static fixed-point filter, while still having uncorrelated, white dither noise.

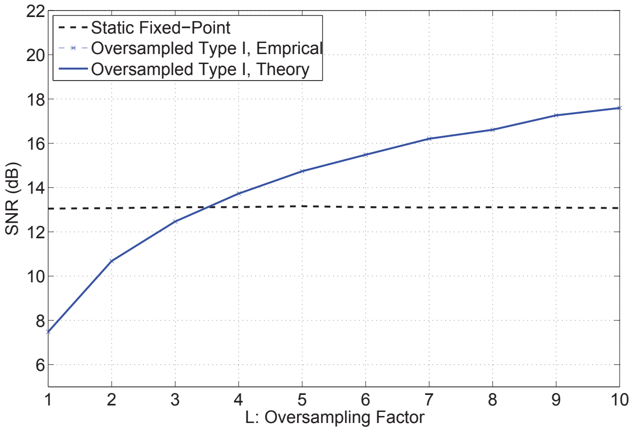


Fig. 3. SNR as a function of oversampling rate ,L, for one-bit Type I dithered implementation of the example of Section 6.

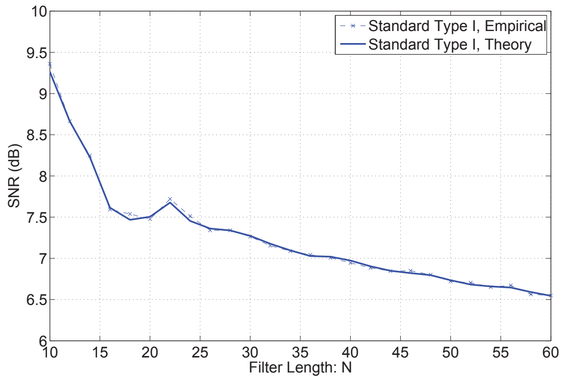


Fig. 4. SNR as a function of N for one-bit standard Type I dithered implementation of the example of Section 6.

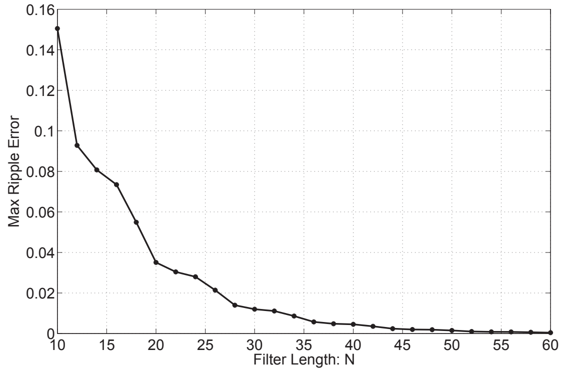


Fig. 5. Max ripple error as a function of N for continuous-valued implementation of the example of Section 6.

7. REFERENCES

[1] R. Greenfield, "Increased precision digital filter coefficients using digital dither," *IEEE Electronics Letters*, vol. 24, no. 5, March 1998.

[2] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1999.

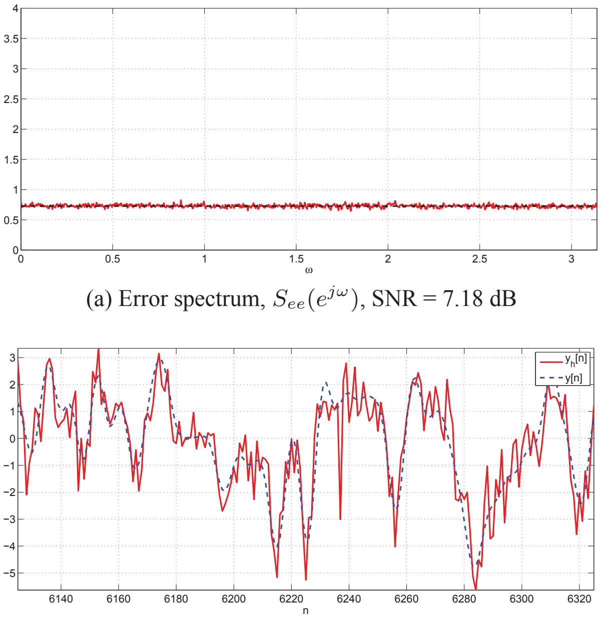
[3] C. Lim, R. Yang, D. Li, and J. Song, "Signed power-of-two term allocation scheme for the design of digital filters," *IEEE Transactions in Circuites and Systems II*, vol. 46, no. 5, pp. 577–584, May 1999.

[4] Jacek Izydorczyk, "An algorithm for optimal terms allocation for fixed point coefficients of FIR filters," in *Proceedings of ICSAS'06*, 2006.

[5] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 155–162, 1981.

[6] M. Said and A.V. Oppenheim, "Discrete-time randomized sampling," in *Proceedings of ICECS'01*. ICECS, Sept 2001.

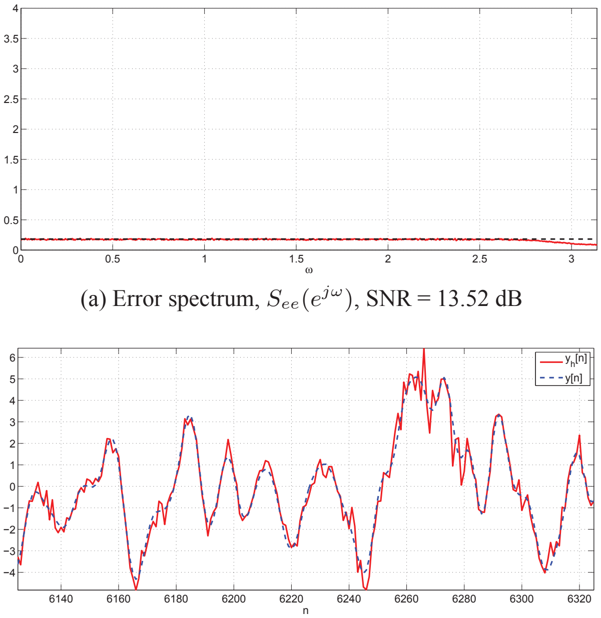
[7] Sourav R. Dey, *Randomized Sampling and Multiplier-Less Filtering*, Ph.D. thesis, Massachusetts Institute of Technology, 2008.



(a) Error spectrum, $S_{ee}(e^{j\omega})$, SNR = 7.18 dB

(b) Time-Domain output $\hat{y}[n]$ and $y[n]$

Fig. 6. Error power spectrum and time-domain output for the one-bit Type I standard dithered implementation of the example of Section 6.



(a) Error spectrum, $S_{ee}(e^{j\omega})$, SNR = 13.52 dB

(b) Time-Domain output $\hat{y}[n]$ and $y[n]$

Fig. 7. Error power spectrum and time-domain output for the one-bit Type I oversampled dithered implementation of the example of Section 6. $L = 4$.