Design and Analysis of Delta-Sigma Based IIR Filters

David A. Johns and David M. Lewis, *Member, IEEE*

Abstract—**This paper presents design techniques for IIR filters operating on oversampled delta–sigma ($\Delta\Sigma$) modulated signals. It is shown that $\Delta\Sigma$-based IIR filters can be efficiently realized by eliminating all multibit multipliers through the use of re-modulating internal filter states. As well, noise results are presented showing that linear noise analysis gives excellent predictions of the noise performance over the frequency band of interest. Finally, it is shown that latency and computational complexity can be reduced in some VLSI applications where digital representations of analog signals exist using oversampled $\Delta\Sigma$ converters.**

I. Introduction

THE USE of oversampled $\Delta\Sigma$ modulation is rapidly gaining popularity as an effective method for building high resolution analog-to-digital (A/D) and digital-to-analog (D/A) converters [1]–[6]. While oversampled converters usually interface to digital signals at the Nyquist rate, it is useful to consider signal processing directly at the oversampled rate in an attempt to save valuable silicon area. For example, in an application where both input and output signals are $\Delta\Sigma$ modulated corresponding to external analog signals, filtering at the Nyquist rate introduces two complementary filters—a decimation and an interpolation filter. Filtering directly on the oversampled signal would eliminate the need for both these filters and could reduce circuit complexity if the $\Delta\Sigma$ based filter is efficiently realized.

One technique for processing $\Delta\Sigma$ modulated signals is to make use of finite-impulse-response (FIR) filters [7]. However, an FIR approach often leads to an excessive number of delay stages (though only one-bit each) and additions when the oversampling ratio is high and the filter order large. Since infinite-impulse-response (IIR) filters can often meet the same specifications with a lower order filter order than their FIR counterparts [8], it is desirable to find efficient $\Delta\Sigma$ based IIR filter realizations. In related work, a method for realizing IIR filters operating on *delta* modulated signals has been proposed [9]; however, since most oversampled A/D and D/A converters are realized using $\Delta\Sigma$ modulation, there is a strong motivation to extend that work.

In an effort to find efficient $\Delta\Sigma$ based IIR filters, a recent publication briefly described a design approach where internal filter states are remodulated using fully digital $\Delta\Sigma$ modulators [10]. This paper expands on the idea presented in [10] by presenting design techniques and a noise analysis. In Section II, the basic approach to $\Delta\Sigma$ based IIR filters is described using a first-order example where it is shown that filter structure is extremely important. Two approaches for realizing higher order transfer-functions based on biquad and quasi-orthonormal structures are presented in Section III with design examples given in Section IV. In Section V, it is shown that although nonlinear modulators are utilized, a linear noise analysis gives excellent agreement with simulation over the frequency band of interest. The use of multibit quantizers are suggested in Section VI and stability for this type of filtering is discussed in Section VII. Finally, a couple of application examples are given in Section VIII where it is shown that reductions in latency and computational complexity are possible.

Before proceeding, some terms relating to oversampling need to be defined. Throughout this paper, we shall assume the frequency of interest to be from 0 to f_o. The oversampling ratio, OSR, is defined to be the ratio of the sampling frequency, f_s, to the Nyquist frequency, $2f_o$. Specifically,

$$\text{OSR} \equiv \frac{f_s}{2f_o}. \tag{1}$$

For simplicity, throughout this paper we shall normalize f_s to unity.

II. Delta-Sigma Based IIR Filtering

The essential building block that allows the construction of $\Delta\Sigma$ based IIR filters is the attenuator circuit shown in Fig. 1(a). A one-bit input signal, $\hat{u}(n)$, operating at the oversampled rate is multiplied by a multibit constant coefficient, a_1, and the resulting multibit signal, $y(n)$, is applied to a digital $\Delta\Sigma$ modulator giving the output one-bit signal, $\hat{y}(n-1)$. Here and throughout this paper, an assumption is made that the $\Delta\Sigma$ modulator introduces a unit delay at f_s from input to output as is the case in many modulators. Examples of signal spectra are shown in Fig. 1(b) where the rising spectra above f_o is due to the modulation noise introduced by the modulator. Modeling the modulator as a single delay plus an additive noise source, $e(n)$, as shown in Fig. 1(c), this circuit clearly behaves as an attenuator over the frequency band of interest with some additional noise. If the common (though sometimes unjustified) assumption is made that noise arising from the quantizer in the modulator is white over the frequency band of interest, then the spectral density of $e(n)$ will be white noise *shaped* by the noise transfer function of the modulator.

Manuscript received April 24, 1992; revised November 17, 1992. This paper was recommended by Associate Editor B. S. Song.

The authors are with the Department of Electrical Engineering, University of Toronto, Toronto, Canada.

IEEE Log Number 9208316.

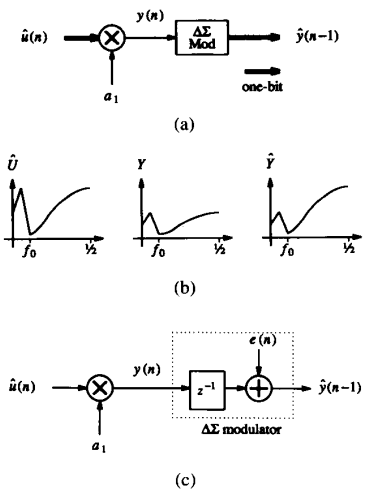


Fig. 1. Delta–sigma attenuator. (a) Circuit implementation. (b) Signal spectra. (c) Equivalent model over the frequency band of interest. The signal $e(n)$ represents the shaped quantization noise introduced by the modulator.

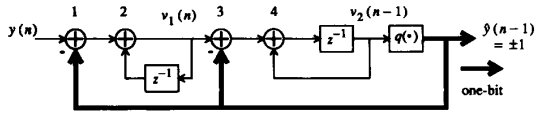


Fig. 2. A digital second-order delta–sigma modulator. The block $q(\cdot)$ denotes a one-bit quantizer implemented by taking the most significant bit.

This structure leads to low hardware complexity since $\hat{u}(n)$ is a one-bit signal and thus the multiplier can be efficiently realized as a 2-input multiplexor with $\hat{u}(n)$ acting as a control input selecting either a_1 or $-a_1$ (simply changing the sign-bit is not sufficient as 2's-complement arithmetic is assumed in order to realize efficient adders). This multiplexor approach results in the $1 \times k$ multiplier requiring about $4k$ transistors in CMOS technology for a k-bit coefficient. For comparison, a $k \times k$ multiplier requires about $28k^2$ transistors. All multipliers shown in this paper are of this multiplexor type—a critical requirement since they must operate at the oversampled rate. The hardware complexity of the modulator depends on the modulator order and architecture. Although many modulators are possible, examples and results throughout this paper will be based on the second-order modulator shown in Fig. 2 [11]. Here, the input and output signals are multibit and one-bit, respectively, and assuming the use of 2's complement arithmetic, the implementation of this modulator can be simplified to contain only two adders. This low complexity is obtained by recognizing the fact that since the value added is either the positive or negative maximum value, the first and third adders require only the addition of a 1-bit signal to the two most significant bits of the k-bit signal. The end result is that the first and third adders can be implemented using only two logic gates each.

For a noise analysis, we make the common (though sometimes unwarranted) approximation that the quantizer is modeled as linearly adding white quantization noise. Specifically, the quantizer in Fig. 2 is replaced by

$$\hat{y}(n-1) = v_2(n-1) + n_q(n) \tag{2}$$

where $n_q(n)$ is the equivalent quantization noise. With this linear circuit model, the output signal, $\hat{y}(n-1)$, can be shown in the z- transform domain to be equal to

$$z^{-1}\hat{Y}(z) = z^{-1}Y(z) + \frac{(z-1)^2}{z^2}N_q(z). \tag{3}$$

Thus, the output signal, $z^{-1}\hat{Y}(z)$, is equal to a delayed version of the input signal plus shaped quantization noise as in Fig. 1(c). For this second-order modulator, the noise transfer-function of (3) is seen to be equal to $(z-1)^2/z^2$.

To realize $\Delta\Sigma$ based IIR filters, high-speed $k \times k$ multipliers can be eliminated by using the approach described above for the attenuator circuit. However, care must be taken in choosing a suitable filter structure since oversampled transfer-functions are a natural consequence when using $\Delta\Sigma$ modulation. Specifically, consider two possible realizations for a first-order low-pass filter as shown in Fig. 3. Substituting the modulator model shown in Fig. 1(c), the output signals are found to be

$$X_a(z) = \frac{b_1}{1-(1-a_1)z^{-1}}\hat{U}(z) + \frac{1-a_1}{1-(1-a_1)z^{-1}}E(z) \tag{4}$$

$$X_b(z) = \frac{b_1}{1-(1-a_1)z^{-1}}\hat{U}(z) + \frac{-a_1}{1-(1-a_1)z^{-1}}E(z) \tag{5}$$

where $X_a(z)$ and $X_b(z)$ are multibit states and $E(z)$ is the modulation noise. The equivalent equations for the one-bit states $\hat{X}_a(z)$ and $\hat{X}_b(z)$ are

$$\hat{X}_a(z) = \frac{b_1 z^{-1}}{1-(1-a_1)z^{-1}}\hat{U}(z) + \frac{1}{1-(1-a_1)z^{-1}}E(z) \tag{6}$$

$$\hat{X}_b(z) = \frac{b_1 z^{-1}}{1-(1-a_1)z^{-1}}\hat{U}(z) + \frac{1-z^{-1}}{1-(1-a_1)z^{-1}}E(z). \tag{7}$$

Defining $X(z)/\hat{U}(z)$ to be the signal gain and $X(z)/E(z)$ to be the noise gain for the multibit signals, we see that both structures have the same signal gains but different noise gains.[1] Since for oversampled transfer-functions the pole nears unity and thus the coefficient a_1 is near zero, clearly the structure in Fig. 3(b) will result in a much lower noise gain than that of Fig. 3(a).[2] Similar noise performance is also true for the one-bit signals, $\hat{X}_a(z)$ and $\hat{X}_b(z)$, as seen from (6) and (7). Another interesting result seen from (5) and (7) is that for frequencies above the pole-frequency, the spectral density of the noise in $\hat{X}_b(z)$ approximates $E(z)$ while the noise in $X_b(z)$ has a spectral density approximating $a_1E(z)/(1-z^{-1})$. This result is due to the location of the integrator with respect to the where the modulation noise is injected. Thus, while the noise in the multibit output signal, $X_b(z)$, is reduced one order in noise-shaping in comparison with $\hat{X}_b(z)$, its total noise power is reduced.

[1] Although the realization in Fig. 3(b) has *two* delaying stages, there are redundant states so that only a first-order transfer-function is realized.

[2] In simulations, the structure in Fig. 3(a) did not work at all. Most likely this was due to the excessive noise gain and the fact that the noise model of Fig. 1(c) is an approximation.

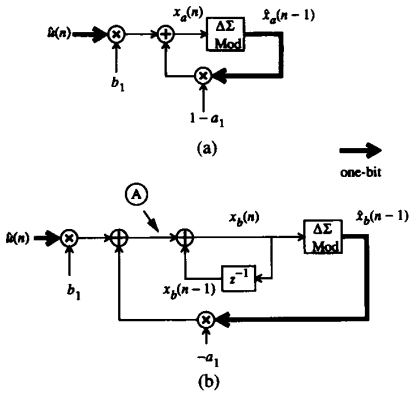


Fig. 3. First-order delta–sigma filter. (a) Direct-form structure with poor noise performance. (b) An integrator-based structure having good noise performance.

Note that since the only noise source in this filter is due to the modulator (the adders are noiseless if no overflow occurs), the dynamic range of this filter will not depend on the number of bits in the adders but rather on the oversampling ratio and the noise shaping ability of the modulator. The number of bits in the adders determined the number of bits representing the coefficients which, in turn, determine the transfer-function accuracy. However, care must be taken to ensure that the internal state $x_b(n)$ does not exceed the modulator's maximum input level and thus dynamic range scaling is important here. Also note that since the $\Delta\Sigma$ modulated signals are binary, the circuit complexity of Fig. 3(b) can be reduced. Specifically, since the signal at node Ⓐ takes on only one of 4 possible values, the adder and two multipliers can be more efficiently realized as a 4-input multiplexor [12]. If the modulator in Fig. 2 is used, this simplification results in the requirement of 3 multibit adders plus some minor logic to realize a first-order filter.

III. Higher Order Filters

Due to the requirement that only single-bit states are multiplied by constant coefficients in $\Delta\Sigma$ based IIR filters, the criteria for choosing a filter structure is different than that of traditional IIR filter design. This section will present two suitable structures where the number of modulators required is equal to the filter order. There are two main reasons for keeping the number of modulators low. One reason is that each modulator may require a significant amount of silicon area, especially if the modulator's order is higher than two. The second reason is to minimize the number of noise sources since, as discussed above, the only sources of noise are in the modulators.[3]

3.1. Biquad Design

Perhaps the most common method of realizing higher order transfer-functions is through the use of a cascade-of-biquads approach. A general biquadratic transfer-function can be written as

$$T(z) = \frac{n_2 z^2 + n_1 z + n_0}{z^2 + p_1 z + p_0}. \tag{8}$$

[3]Of course, one should ensure the resulting filter structure does not have excessive noise gains as in the direct-form structure.

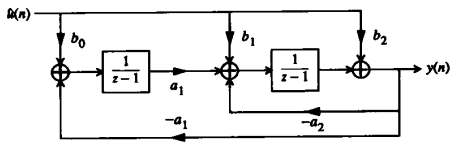


Fig. 4. A general biquad structure using two delaying integrators. Note that input summing is used to form the transfer-function zeros.

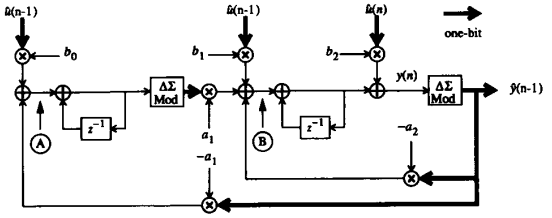


Fig. 5. A general biquad $\Delta\Sigma$-based IIR filter. The output can be either the one-bit signal $\hat{y}(n-1)$ or the multibit signal $y(n)$.

Assuming the modulator introduces a unit delay, the general biquad structure should be based on two *delaying* integrators as was the first-order filter in Fig. 3(b). One such biquad structure is shown in Fig. 4 where input summing is used to obtain the correct transfer-function zeros. While output summing is typically used in biquad designs, input summing is used here to eliminate the need for an extra modulator in forming a one-bit output signal. The symmetrical coefficients, a_1 and $-a_1$, are used to maintain good dynamic range scaling for the internal multibit states. This scaling is important in oversampled functions and is accomplished by making the integrators have equal time-constants, similar to the approach used in continuous-time filtering.

After some signal-flow-graph manipulations, a biquad $\Delta\Sigma$-based IIR filter can be obtained based on the structure in Fig. 4 as shown in Fig. 5. The *delayed* input signals are a result of moving the summing node associated with the b_2 coefficient to the other side of a delay stage. Fortunately, delaying the signal $\hat{u}(n)$ requires a trivial amount of extra circuitry since $\hat{u}(n)$ is a one-bit signal. The transfer-function for the filter in Fig. 5 can be shown to be given by

$$T(z) \equiv \frac{Y(z)}{\hat{U}(z)} = \frac{b_2 z^2 + (b_1 - 2b_2)z + (b_0 a_1 - b_1 + b_2)}{z^2 - (2 - a_2)z + (1 + a_1^2 - a_2)}. \tag{9}$$

Equating the coefficients in (9) with those in (8) results in the following design equations:

$$a_2 = p_1 + 2 \tag{10}$$

$$a_1 = \sqrt{p_0 + p_1 + 1} \tag{11}$$

$$b_2 = n_2 \tag{12}$$

$$b_1 = n_1 + 2n_2 \tag{13}$$

$$b_0 = \frac{1}{a_1}(n_0 + n_1 + n_2). \tag{14}$$

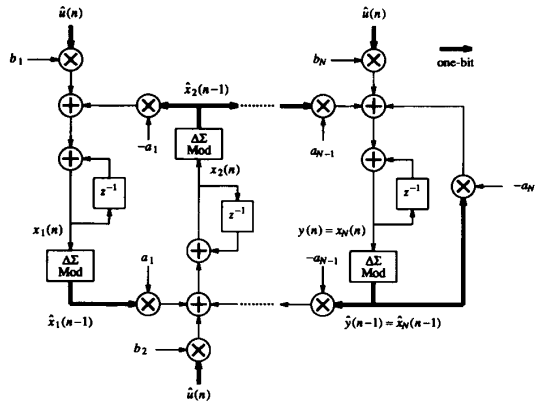


Fig. 6. An Nth-order delta–sigma IIR filter using the quasi-orthonormal structure.

Recognizing that the signals at nodes Ⓐ and Ⓑ take on only one of either 4 or 8 values, respectively, multiplexors can again be used to reduce the hardware complexity of this general biquad to only 7 multibit adders together with some minor logic assuming the modulator shown in Fig. 2 is used. Thus, in general, an Nth-order filter (if N is even) would require $3.5N$ adders.

3.2. *Quasi-Orthonormal Design*

An alternate structure for realizing oversampled transfer-functions with good noise and sensitivity performance is to use a quasi-orthonormal state-space structure [13], [14]. For high oversampling ratios, the quasi-orthonormal structure has two main advantages. One advantage is that the structure is inherently scaled for optimum dynamic range in terms of an L_2 norm making it extremely useful for programmable applications. Secondly, it has a unique representation for a given transfer-function (within scaling factors of ± 1) and a performance comparable to an optimum biquad cascade design where pole-zero pairing and cascade ordering need to have been carefully chosen. Making use of this structure[4] results in the Nth-order IIR filter shown in Fig. 6. Each multibit state signal, $x_i(n)$, is applied to a $\Delta\Sigma$ modulator resulting in a one-bit signal $\hat{x}_i(n-1)$.

Note that making use of 8-input multiplexors as before, the hardware complexity of this structure requires $3N$ multibit adders, which is slightly less than that for a cascade-of-biquads design.

IV. Design Examples

Simulations were performed for a number of different filters, structures, and modulators. For the sake of brevity, results are described here for only an eighth-order bandpass and three fifth-order low-pass filters using the quasi-orthonormal structure with the modulator depicted in Fig. 2. The bandpass filter had an oversampling ratio of 128 while the fifth-order filters were all based on a single prototype filter but frequency scaled to three different oversampling ratios; $OSR = 32$, 64, and 128. The coefficient values for these filters are given in Table I where it is clear that the coefficients approach zero as the oversampling ratio is increased.

[4]In fact, the transposed structure was used which transforms an output summing stage to the shown input summing stage consisting of the b coefficients.

TABLE I
Coefficient Values for Quasi- Orthonormal $\Delta\Sigma$ Filters. OSR is the Oversampling Ratio

Coeff	Low-pass			Bandpass
	OSR = 32	OSR = 64	OSR = 128	OSR = 128
a_1	0.0221814	0.0113349	0.0057310	0.0092571
a_2	0.0161338	0.0079576	0.0039475	0.0017366
a_3	0.0186953	0.0092801	0.0046218	0.0101197
a_4	0.0259125	0.0127442	0.0063178	0.0021232
a_5	0.0304938	0.0149179	0.0073764	0.0098791
a_6	—	—	—	0.0031290
a_7	—	—	—	0.0104593
a_8	—	—	—	0.0045269
b_1	0.0221042	0.0111391	0.0055940	0.0000390
b_2	0.0010365	0.0002694	0.0000688	−0.0026590
b_3	0.0056275	0.0028927	0.0014681	−0.0000388
b_4	0.0001224	0.0000313	0.0000079	0.0010142
b_5	0.0004097	0.0002061	0.0001034	0.0000110
b_6	—	—	—	0.0000619
b_7	—	—	—	−0.0000011
b_8		—	—	−0.0000588

To measure the frequency responses, sine waves at varying frequencies were first passed through a $\Delta\Sigma$ modulator, then the filter, and finally an FFT and Hanning window of the resulting one-bit stream indicated the magnitude response at the sine wave's frequency. Although this method of measurement is much more computationally intensive than using an impulse response (as in [7]), it allows more power at each frequency and therefore gives a better estimate of the frequency response for a given oversampling ratio. In Fig. 7, the frequency responses for the four filters are shown along each individual signal-to-noise ratio (SNR). The SNR was determined by the ratio of the output power of a single sinusoid (peak values of $\pm\frac{1}{4}$ for quantizer output levels of ± 1) at the upper passband edge of the filter to the total output noise power over the frequency of interest. Here, we see that the signal-to-noise performance improves for higher oversampling ratios, as expected.

It is of interest to compare the dynamic range of these filters to that of a single $\Delta\Sigma$ modulator. The simulated values of SNR for the signal $\hat{u}(n)$ for oversampling ratios of 32, 64 and 128 were 51, 65, and 80, respectively. Comparing these values with the SNR values of Fig. 7, we see that about a loss of about 1 dB occurs in the fifth-order filters while a loss of 5 dB results in the higher-Q eighth-order filter.

V. Noise Performance

Since $\Delta\Sigma$-based IIR filtering relies on re-modulating internal filter states where the states have significant modulation noise introduced from other modulators, it is important to investigate the noise behavior of the resulting filters. Specifically, it will be shown in this section that a simple linear

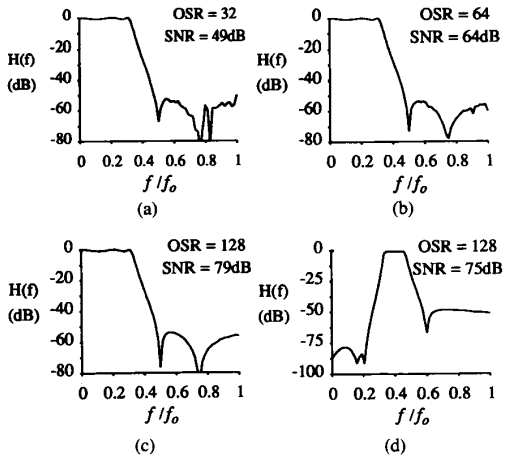


Fig. 7. Frequency responses for various $\Delta\Sigma$-based IIR filters. (a)–(c) fifth-order lowpass; (d) eighth-order bandpass.

noise analysis is successful in accurately predicting the noise performance for these types of filters over the frequency band of interest.

A linear noise model for a $\Delta\Sigma$-based filter can be obtained by replacing each of the $\Delta\Sigma$ modulators with a delay stage plus an additive noise source, $e_i(n)$, as was done in Fig. 1. As well, one extra noise source, $e_{N+1}(n)$, can be used to model the noise added by a modulator operating on the input signal to create the one-bit signal, $\hat{u}(n)$. The noise spectral density, $S_e(f)$, of $e(n)$ is determined by the modulator choice.

Defining $W_i(f)$ to be the noise gain from $e_i(n)$ to the output signal and assuming each of the noise sources, $e_i(n)$, to be uncorrelated, the total equivalent noise gain, $W(f)$, is defined as

$$W(f) = \left(\sum_{i=1}^{N+1} |W_i(f)|^2 \right)^{1/2}. \tag{15}$$

Finally, the noise spectral density of the output signal, $S(f)$, is easily shown to be given by

$$S(f) = |W(f)| S_e(f). \tag{16}$$

As an example, consider the fifth-order filter with $OSR = 32$ described in Section IV where the second-order modulator shown in Fig. 2 was used. Modeling the quantizer as adding noise uniformly distributed between in $[-1, +1]$, the mean-square noise value equals $\frac{1}{3}$. Assuming a two-sided representation of frequencies for spectral density calculations, all the noise power of a quantized signal sampled at frequency $f_s = 1$ is folded into the frequency band $(-1/2) \leq f < (1/2)$. If the quantization noise is also white, then the spectral density of the quantization noise is $\frac{1}{3}$. This quantization noise is shaped by the noise transfer-function, $(1 - z^{-1})^2$, resulting in the spectral density, $S_e(f)$, of $e(n)$ for this second-order modulator given by (as in [1])

$$\begin{aligned} S_e(f) &= \frac{1}{\sqrt{3}} |1 - e^{-j2\pi f}|^2 \\ &= \frac{4}{\sqrt{3}} \sin^2 \left(\frac{2\pi f}{2} \right). \end{aligned} \tag{17}$$

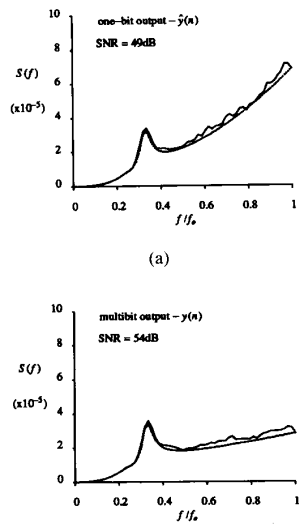


Fig. 8. Simulated versus predicted in-band noise spectral densities for the output signals, $\hat{y}(n)$ and $y(n)$ for a fifth-order filter.

The expected noise spectral density of the output is then determined from (16) using $S_e(f)$ found in (17) and $W(f)$ found from transfer-function analysis applied to the fifth-order filter. A comparison between the expected and simulated noise curves is presented in Fig. 8(a) and (b). The simulated noise curves were obtained by averaging 256 periodogram estimates each of length 4096 resulting in 64 FFT bins over the frequency band of interest. Note that the output signal can be considered to be either the one-bit signal, $\hat{y}(n)$, or the multibit signal, $y(n)$. In-band noise comparisons were also performed for the eighth-order bandpass filter and gave similar agreement. This excellent agreement indicates that the simple model of Fig. 1(c) is valid for predicting the noise performance over the frequency band of interest.

While in-band noise for these two output signals is approximately the same (within 5 dB), out-of-band noise is significantly higher (about 30 dB higher) for $\hat{y}(n)$ as shown in Fig. 9(a) and (b). Both the in-band and out-of-band noise on $y(n)$ are lower due to all the one-bit modulated signals being filtered by the last integrator that forms $y(n)$. This filtering effect is clearly seen in Fig. 8 for frequencies above the passband edge where the noise on $y(n)$ rises slower than that for $\hat{y}(n)$. The in-band noise difference of 5 dB is much smaller than the out-of-band noise difference of 30 dB simply due to the fact that the noise is integrated over a smaller frequency band. Clearly, if the output of this filter were to go into a decimation stage, the output of choice would be the multibit signal, $y(n)$, where out-of-band noise has already been reduced.

VI. The Use of Multibit Quantizers

This paper has thus far assumed that the $\Delta\Sigma$ modulator uses a two-level quantizer. The advantage of a two level quantizer is that multiplication by ± 1 is easily accomplished

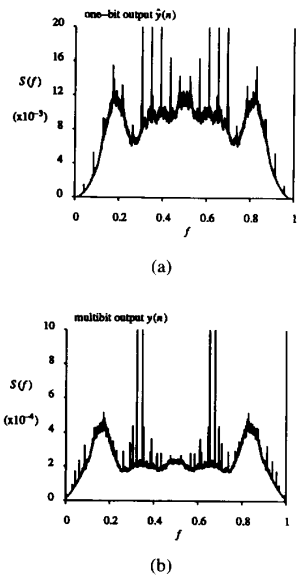


Fig. 9. Out-of-band noise spectral densities for the output signals, $\hat{y}(n)$ and $y(n)$ for fifth-order example. Note that the noise for $y(n)$ is much lower than that for $\hat{y}(n)$.

with a multiplexer. However, it is useful to consider multibit quantizers that also have simple implementations for a couple of reasons. First, it is well known that quantization noise is proportional to quantization step size and thus dynamic range is increased by 6 dB as the number of quantization levels is doubled. The second reason is that multibit quantizers tend to follow the predictions of linear theory more closely than their one-bit counterparts and are thus more likely to be stable [1]. It should be mentioned here that multibit quantizers are often avoided in data converters due to the possibility of nonlinearity. However, perfect linearity can be maintained here since the entire modulator is digital.

Since one of the features of the $\Delta\Sigma$ filter is to allow simple multiplication by a quantized signal, it is instructive to look to multiplier techniques. Booth recoded multipliers [15] recode a multiplier into a smaller number of higher-radix digits. While a Booth recoded multiplier of arbitrary order is possible, the highest order Booth multiplier that allows multiplexer only designs is the second-order Booth multiplier. This second-order multiplier has a redundant radix-4 representation using the digit set $\{-2, -1, 0, 1, 2\}$. Multiplication by any of these digits can be accomplished by a 5-input multiplexer using the multiplicand and its negative, and using a one-bit shifted version for multiplication by ± 2. This suggests the use of a five-level quantizer with output levels 0, ± 0.5, and ± 1 while the threshold levels placed for convenience at $\pm\frac{1}{4}$ and $\pm\frac{3}{4}$. Simulations have shown that a five-level quantizer results in about a 13-dB SNR improvement which is close to the 12-dB expected result due to the reduced quantization size.

It should be noted that the hardware complexity when using 5-level quantizers in the quasi-orthonormal structure becomes $5N$ multibit adders rather than $3N$. This extra complexity is a result of using 2 extra adders to realize the signals entering each integrator rather than 8-input multiplexors generating these signals as before. In summary, for some additional hardware, a five-level quantizer would be almost equivalent to increasing the oversampling rate by two for this second-order modulator.

VII. Stability

So far in this paper, we have modelled $\Delta\Sigma$ modulators as single delay stages plus shaped quantization noise and then used linear theory for both design and analysis. However, an alternate viewpoint is to regard these systems as higher order modulators with potentially multiple quantizers. For example, note that the first-order $\Delta\Sigma$-based IIR filter in Fig. 3(b) can be regarded as a third-order $\Delta\Sigma$ modulator with a first-order signal transfer-function. Thus, the question of stability for this higher order modulator becomes important. Unfortunately, while there have been some rigorous stability analyses for specific modulators [16], [17] and a conservative stability criteria for arbitrary modulators [18], the stability of this third-order modulator is not easily determined. However, some insight can be gained by the looking at the noise transfer-function, NTF, which for the case of the filter of Fig. 3(b) using the modulator of Fig. 2 is given by,

$$\mathrm{NTF} = \frac{(z-1)^3}{z^2(z-(1-a_1))} \tag{18}$$

Note that this NTF has the same noise transfer-function as that of the second-order modulator except that an extra zero and pole occur at +1 and $1 - a_1$, respectively. Since a_1 is positive and close to zero as discussed earlier, it is clear that this third-order NTF is approximately the same as the second-order NTF except near dc. While this observation does not guarantee stability of the third-order system given that the second-order modulator is stable, it does indicate that the rule-of-thumb stability criteria based on the NTF given in [19]–[21] would be maintained. In higher order $\Delta\Sigma$-based IIR filters, multiple quantizers exist resulting in multiple noise transfer-functions and thus rigorous stability criteria may be even more difficult to find. Fortunately, simulations of biquad filters with Q-factors as high as a few thousand indicate that stability of these systems are determined by the stability of the equivalent linear system assuming the modulators themselves are stable. The authors are currently attempting to find either a proof which justifies this conjecture or a rigorous test that certifies the stability of these filters.

VIII. Application Examples

Perhaps the most obvious application for $\Delta\Sigma$-based IIR filtering is to eliminate both the decimation and interpolation filters when applying digital IIR filtering on analog signals as shown in Fig. 10. The approach in figure 10(a) makes use of decimation and interpolation filters such that a near Nyquist rate DSP can be utilized while that of Fig. 10(b) operates solely at the oversampled rate using the technique described in this paper. By eliminating the decimation and interpolation filters we can expect two benefits—a reduced input/output latency and a reduction in circuit complexity

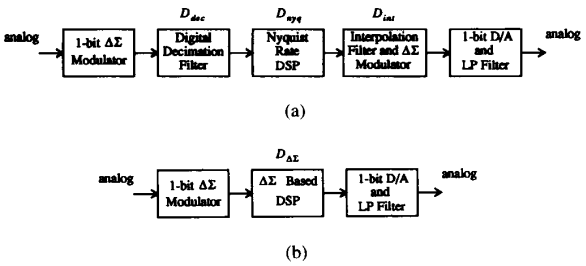


Fig. 10. Alternate DSP realizations for filtering on analog signals. (a) Traditional Nyquist-Rate filtering. (b) Oversampled $\Delta\Sigma$ based filtering. $D_{\rm dec}, D_{\rm nyq}, D_{\rm int}$, and $D_{\Delta\Sigma}$ indicate the number of additions required per sample at the Nyquist rate for the various blocks.

The extra latency in the Nyquist rate filter is a result of the near brickwall decimation and interpolation filters required to reduce aliasing and imaging. The potential complexity reduction is a function of the IIR filter orders being realized. As a rough figure of merit to compare complexity for these two approaches, we define $D_{\rm dec}, D_{\rm nyq}, D_{\rm int}$, and $D_{\Delta\Sigma}$ to be the number of additions required per sample at the Nyquist rate for the various blocks as shown. Although various structures can be used for the Nyquist rate filter, generally $3N\,k \times k$ bit multiplications are required for a good performance Nth-order filter. Thus, the value of $D_{\rm nyq}$ is seen to be approximately $3Nk$ where we assume a $k \times k$ bit multiplication requires k additions. The values of $D_{\rm dec}$ and $D_{\rm int}$ are fixed for a given application and depend on a variety of issues such as implementation choice, phase and noise requirements, etc. Finally, for an Nth-order $\Delta\Sigma$-based filter, $D_{\Delta\Sigma}$ equals $5N \times OSR$ assuming second-order modulators with 5-level quantizers are used. Thus as a rough comparison, the total number of additions/sample for the Nyquist rate approach is given by

$$T_{\rm nyq} = D_{\rm dec} + 3N \times k + D_{\rm int} \tag{19}$$

while the total for the oversampled approach is

$$T_{\Delta\Sigma} = 5N \times OSR. \tag{20}$$

We see from (19) and (20) that $T_{\Delta\Sigma}$ will be greater than $T_{\rm nyq}$ for large values of N since OSR is always larger than k. What is not so clear is the potential complexity trade-off for lower orders of N.

As an example, consider an 18-bit A/D converter with $OSR = 64$ described in [3]. For this converter, a 4096 tap FIR decimation filter was used resulting in $D_{\rm dec} = 4095$. Assuming an interpolation filter of roughly the same complexity as that of the decimation filter, $T_{\rm nyq}$ for the Nyquist rate approach would be given by $8190 + 54N$. For the oversampled approach to meet an 18-bit performance using second-order modulators and 5-level quantizers would require OSR = 128 resulting in $T_{\Delta\Sigma} = 640N$. Comparing (19) and (20) for this example results in the oversampled approach requiring less additions/sample when the filter order, N, is less than 14. As well, the latency due to both the decimation and interpolation filters is 64 samples at the Nyquist rate. In summary, processing directly at the oversampled rate will be advantageous in low to medium orders of filtering and where extra latency cannot be tolerated.

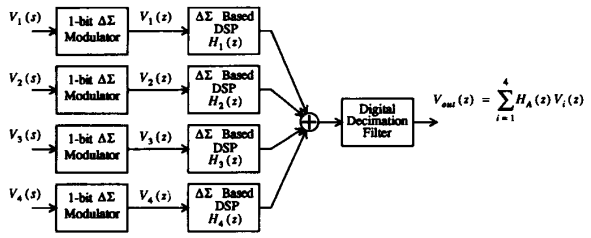


Fig. 11. A multi-input A/D converter where each input requires programmable gain and filtering.

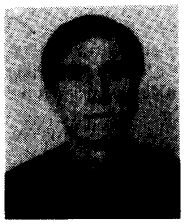
As a second application, consider the multi-input A/D converter shown in Fig. 11 where the inputs are individually filtered and summed together. Examples of this application are wideband beamforming or audio mixing boards. The traditional approach would require four decimation filters whereas only one is required if the necessary signal processing can be performed at the oversampled rate. Using the same numerical values as in the previous example, the oversampled approach would require less additions/sample when the total order for all four filters combined is 20 or less.

IX. Conclusions

Design techniques for realizing IIR filters operating on oversampled $\Delta\Sigma$ modulated signals were presented. It was shown that efficient VLSI realizations can be obtained if multibit multipliers are eliminated by re-modulating internal filter states. A first-order example demonstrated that filter structure is critical since oversampled transfer-functions must be realized and structures were suggested for higher order filters. In terms of hardware complexity, it was shown that only $3N$ adders together with some minor logic are required to implement an Nth-order filter using second-order modulators. Through examples, it was seen that the noise performance for $\Delta\Sigma$-based IIR filters resulted in a small reduction of dynamic range over a standard modulator. Noise analysis and simulated results were also presented showing that a linear noise analysis gives an excellent prediction of noise performance over the frequency band of interest. It was also suggested that with some extra hardware, a five-level quantizer would result in an extra 13 dB of dynamic range. Stability was then discussed where although linear analysis again appears to be sufficient, clearly a more rigorous criteria or justification is needed. Finally, example applications were presented demonstrating that this type of filtering should prove useful in VLSI technologies where interfaces to analog signals are required.

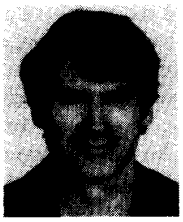
[1] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," *Oversampling Delta–Sigma Converters*, J. C. Candy and G. C. Temes, Eds. IEEE Press, 1992.
[2] B. P. Signore, D. A. Kerth, N. S. Sooch, and E. J. Swanson, "A monolithic 20-b delta-sigma A/D converter," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1311–1317, Dec. 1990.
[3] P. Ferguson, Jr. *et al.*, "An 18b 20kHz dual sigma-delta A/D converter," *ISSCC91 Digest of Tech. Papers*, vol. 34, Feb. 1991.

[4] B. P. Brandt and B. A. Wooley, "A 50-MHz multibit sigma-delta modulator for 12-b 2-MHz A/D conversion," *IEEE J. Solid State Circuits,* vol. 26, pp. 1746–1756, Dec. 1991.
[5] Y. Matsuya, K, Uchimura, A. Iwata, and T. Kaneko, "A 17-bit oversampling D-to-A conversion technology using multistage noise shaping," *IEEE J. Solid-State Circuits,* vol. 24, pp. 969–975, Aug. 1989.
[6] H. J. Schouwenaars, D. W. J. Groeneveld, C. A. A. Bastiaansen, and H. A. H. Termeer, "An oversampled multibit CMOS D/A converter for digital audio with 115-dB dynamic range," *IEEE J. Solid-State Circuits,* vol. 26, pp. 1775–1780, Dec. 1991.
[7] P. W. Wong and R. M. Gray, "FIR Filters with delta-sigma modulation encoding," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 38, pp. 979–990, June 1990.
[8] B. Beliczynski, I. Kale, and G. D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction," *IEEE Trans. on Signal Processing,* vol. 40, pp. 532–542, Mar. 1992.
[9] H. Padir and L. E. Franks, "A new digital recursive filter structure based on delta modulation encoding," in *Proc. ICASSP,* pp. 1850–1853, Apr. 1988.
[10] D. A. Johns and D. M. Lewis, "IIR filtering on delta-sigma modulated signals," *Electron. Lett.,* vol. 27, pp. 307–308, Feb. 1991.
[11] J. C. Candy, "A use of double integration in sigma-delta modulation," *IEEE Trans. Commun.,* vol. 33, pp. 249–258, Mar. 1985.
[12] A. Peled and B. Liu, "A new approach to the realization of nonrecursive digital filters," *IEEE Trans. Audio Electroacoust.,* vol. AU-21, pp. 477–484, Dec. 1973.
[13] D. A. Johns, W. M. Snelgrove, and A. S. Sedra, "Adaptive recursive state-space filters using a gradient-based algorithm," *IEEE Trans. Circuits Syst.,* vol. 37, pp. 673–684, June 1990.
[14] ——, "Orthonormal ladder filters," *IEEE Trans. Circuits Syst.,* vol. 36, pp. 337–343, Mar. 1989.
[15] H. Sam and A. Gupta, "A generalized multibit recoding of two's complement binary numbers and its proof with application in multiplier implementations," *IEEE Trans. Computers,* vol. 39, pp. 1006–1015, Aug. 1990.
[16] S. Hein and A. Zakhor, "On the stability of interpolative sigma delta modulators," in *Proc. ISCAS'91,* pp. 1621–1624, June 1991.
[17] H. Wang, "A geometric view of $\Sigma\Delta$ modulations," *IEEE Trans. Circuits Syst.—II,* vol. 39, pp. 402–405, June 1992.
[18] D. Anastassiou, "Error diffusion coding for A/D conversion," *IEEE Trans. Circuits Syst.,* vol. 36, pp. 1175–1186, Sept. 1989.
[19] B. P. Agrawal and K. Shenoi, "Design methodology of $\Sigma\Delta M$," *IEEE Trans. Commun.,* vol. 31, pp. 360–370, Mar. 1983.
[20] K. C.-H. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini, "A higher order topology for interpolative modulators for oversampling A/D converters," *IEEE Trans. Circuits Syst.,* vol. 37, pp. 309–318, Mar. 1990.
[21] T. Ritoniemi, T. Karema, and H. Tenhunen, "Design of stable high order 1-bit sigma-delta modulators," in *Proc. ISCAS'90,* pp. 3267–3270, May 1990.



David A. Johns received the B.A.Sc., M.A.Sc. and Ph.D. degrees from the University of Toronto, Canada, in 1980, 1983 and 1989, respectively.

From 1980 to 1981 he worked as an applications engineer in the semiconductor division of Mitel Corp., Ottawa, Canada. From 1983 to 1985 he was an analog IC designer at Pacific Microcircuits Ltd., Vancouver, Canada. Upon completion of his doctoral work, he was hired at the University of Toronto where he is currently an assistant professor. His research interests are in the areas of integrated circuit design and signal processing.



David Lewis (M'87) received the B.A.Sc. degree with honors in engineering science from the University of Toronto in 1977, and the Ph.D. degree in electrical engineering in 1985.

From 1982 to 1985 he was employed as a research associate on the Hubnet project, and developed custom integrated circuits for a 50Mb/s local area network. He has been an assistant professor at the University of Toronto since 1985, and associate professor since 1991. His research interests include logic and circuit simulation, logarithmic arithmetic, signal processing field programmable hardware, and VLSI architecture.

Dr. Lewis is a member of the ACM.