

# Efficient Single-Bit Ternary Digital Filtering Using Sigma-Delta Modulator

Adam Charles Thompson, *Student Member, IEEE*, Peter O'Shea, Zahir M. Hussain, *Member, IEEE*, and Brenton R. Steele, *Student Member, IEEE*

**Abstract**—Efficient filtering of sigma-delta bit-streams using a finite-impulse response (FIR)-like digital filter is presented. The filter combines a lowpass sigma-delta modulator system and a ternary FIR filter. Unlike conventional FIR filters, the combination of the two components allows a bit stream to be filtered, with the output being retained in single-bit format.

**Index Terms**—Digital filtering, finite-impulse response (FIR), infinite impulse response (IIR), modulation, sigma-delta, single-bit, ternary filter.

## I. INTRODUCTION

THE SHORT word-length, often single bit, format generated by sigma-delta modulators ( $\Sigma\Delta M$ ) often makes for greatly simplified arithmetic processing. For hardware implementation, this simplified processing implies reduced silicon space. Processing tasks that are rich in multiplications are particularly strong beneficiaries of the use of single-bit signal representations. This is so because, multibit multiplications require complex circuit implementations containing very large number of transistors. However, in the single-bit domain, multiplications can simply be implemented using a multiplexer.

Recently a number of techniques for single-bit processing of  $\Sigma\Delta M$  single-bit streams have been presented [1]–[3]. In [3], the author makes use of a fourth-order  $\Sigma\Delta M$  and a zero-interleaved multibit finite-impulse response (FIR) filter. However, this is not as efficient as the  $\Sigma\Delta M$ -based IIR filter in [1]. The latter technique needs only multiplexers, without the parallel multibit multipliers that are required by the former technique. On the other hand, the IIR-based filter suffers from the disadvantages that the phase is no longer linear, and that the filter is much more vulnerable to coefficient quantization errors than standard FIR filters. To alleviate problems due to the IIR filter coefficient quantization it is proposed in [1] that higher order IIR filters be implemented with quasi-orthonormal structures. These structures require  $N$   $\Sigma\Delta M$ s if an  $N$ th-order IIR filter is to be realized. This proliferation of  $\Sigma\Delta M$ s greatly reduces the implementation efficiency. In addition, increasing the number of

Manuscript received March 21, 2003; revised May 14, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xi Zhang.

A. C. Thompson, Z. M. Hussain, and B. R. Steele are with the School of Electrical and Computer Engineering, RMIT, Melbourne, Victoria 3000, Australia (e-mail: s2114313@student.rmit.edu.au; zmhussain@ieee.org; bsteele@dynamichearing.com.au).

P. O'Shea is with the School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Brisbane, Qld. 4000, Australia (e-mail: pj.oshea@qut.edu.au).

Digital Object Identifier 10.1109/LSP.2003.821734

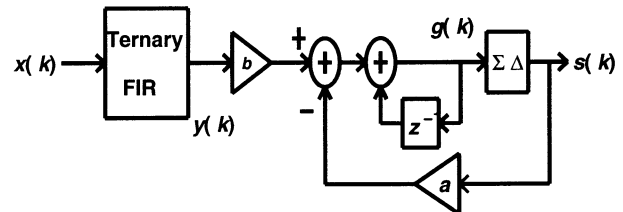


Fig. 1. Block diagram of the proposed digital  $\Sigma\Delta$  FIR-like bit-stream filter.

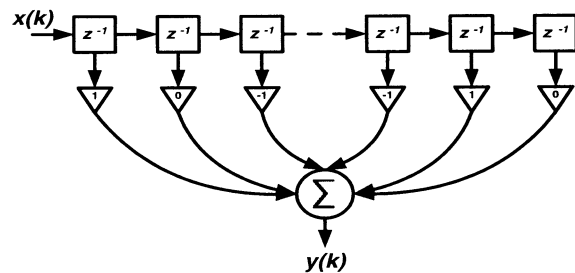


Fig. 2. Block diagram of a ternary FIR filter.

modulators adds to the in-band noise in these structures because, the modulators are the main source of noise in these filters. In this letter, we propose the use of a FIR-like  $\Sigma\Delta$  bit-stream filter. The filter consists of two main sections, a ternary filter and a recursive remodulating filter. The structure of the filter is shown in Fig. 1. The values of the coefficients  $a$  and  $b$  would be determined in Section III.

## II. TERNARY FIR FILTER

The ternary filter is an FIR filter with ternary taps (i.e.,  $+1$ ,  $-1$ , and  $0$ ). The ternary nature of the taps allows a simple implementation of the FIR filter. This filter is extremely efficient when the input signal to the filter is in single-bit format; each multiplication in the FIR filtering operation can be then implemented in hardware with either a couple of logic gates or a very simple lookup table [5]. The structure of the ternary filter is shown in Fig. 2.

Mathematically, the FIR filter output  $y(k)$  can be described by a convolution of the ternary taps  $h_i$  and the input signal  $x(k)$ . If  $M$  is the order of the filter the output of the filter is

$$y(k) = \sum_{i=0}^M h_i x_{k-i}, \quad h_i \in \{1, 0, -1\}. \quad (1)$$

The  $\Sigma\Delta$  filter described here is able to accept both multibit and single-bit input words. The use of multibit words will of

course add to the complexity to the final implementation. The zero-valued taps in the ternary encoded impulse response enhances the efficiency of implementation. For these tap values, no multiplications or additions are required, only delays. This reduces the maximum number of bits that is required for the multibit summer within the ternary FIR filter. We have found, that for simple lowpass filters the number of zero-valued taps could be as high as half of the total number of taps.

The tap values are generated via either  $\Sigma\Delta$  modulation of a target impulse response, or by using optimization techniques discussed in [4] and [5]. For the purpose of this letter, it will be assumed that  $\Sigma\Delta$  modulation is used. Before a target impulse response is encoded to a ternary format, it must be scaled so that the maximum input to the  $\Sigma\Delta M$  is operating at its maximum signal-to-quantization-noise ratio. This scaling produces a magnification of the input signal, but this magnification can easily be removed later, as discussed in Section III.

The digital  $\Sigma\Delta M$  used to generate the ternary filter taps must meet two criteria. Firstly, a ternary quantizer is required to generate the trilevel output; this has the advantage of higher SNR than the common single-bit quantizer [2]. Based on results recorded in [2], we chose equally spaced thresholds about the ternary levels  $\{-1, 0, 1\}$ . The second criteria is that the  $\Sigma\Delta M$  have a flat signal frequency response over the bandwidth  $f_o$  of the signal, i.e., the  $\Sigma\Delta M$  should not unduly modify the shape of the impulse response; it should only add quantization noise, which is largely confined to the out-of-band region. The ternary filter requires operation at an oversampled rate (OSR), a requirement that will be met since the input signal is assumed to be a  $\Sigma\Delta$ -modulated bit-stream. Several  $\Sigma\Delta M$  structures and orders that satisfies the above conditions have been tested. We found little difference between these structures and orders. However, this matter should be further investigated in future works. In this letter, we utilized a second order  $\Sigma\Delta M$  whose structure is shown in Fig. 3.

The  $z$  domain transfer function of the  $\Sigma\Delta M$  shown in Fig. 3 is given by [7]

$$H(z) = G(z)z^{-1} + E(z)(1 - 2z^{-1} + z^{-2}) \quad (2)$$

where  $G(z)$  represents the target impulse response and  $E(z)$  represents the quantization noise transfer functions. The noise-shaping effect of the  $\Sigma\Delta M$  is evident from the presence of the filtering term  $(1 - 2z^{-1} + z^{-2})$ , acting on the noise term  $E(z)$ . The frequency response of the above  $\Sigma\Delta M$  is given by

$$H_{\Sigma\Delta T}(e^{j\Omega}) = G(e^{j\Omega})e^{-j\Omega} + E(e^{j\Omega})(1 - 2e^{-j\Omega} + e^{-2j\Omega}) \quad (3)$$

where  $\Omega$  is the normalized radian frequency.

One advantage of a low-bit resolution system is that the coefficient quantization noise falls in the same spectral region outside  $f_o$  as the input signal quantization noise and the remodulating filter quantization noise.

### III. IIR $\Sigma\Delta$ REMODULATING FILTER

The ternary FIR filter suffers from two disadvantages. It still contains some high-frequency noise due to the coarse quantization of both the impulse response and the input signal. Second

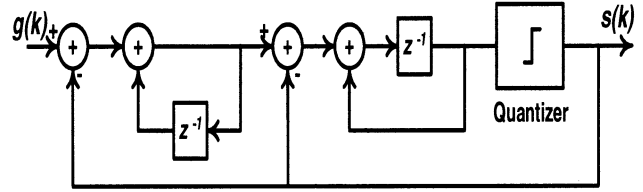


Fig. 3. Block diagram of the second-order  $\Sigma\Delta$  modulator.

it also produces a multibit output. Such outputs are not as conducive to efficient hardware processing as single-bit output. To put the output in single-bit format, and to reduce some of the high-frequency noise, a recursive remodulating filter is used as shown in Fig. 1. The tasks of remodulating the output of the ternary filter and reducing the high-frequency noise cannot be achieved efficiently by using conventional digital  $\Sigma\Delta M$ . These modulators generally have an allpass signal frequency response; hence, these are vulnerable to stability problems caused by high-frequency components at the input. This is due to the fact that the high-frequency energy components increase the quantizers input variance. As a result, the AC loop gain and stability margin are reduced [3]. To overcome this difficulty, a structure with a lowpass signal transfer function and a single-bit output are required.

In [1], several remodulating structures are proposed. These structures contain an IIR filter with embedded  $\Sigma\Delta M$ . The simplest form of such a recursive filter has a first-order IIR structure. The digital  $\Sigma\Delta M$  used in this filter must only introduce a single delay throughout the system. This arises because the  $\Sigma\Delta M$  is used as a delay element in the IIR filter, and as such this limits the selection of  $\Sigma\Delta M$ 's. The best choice of  $\Sigma\Delta M$  should provide good noise shaping at low OSRs. The requirement for a relatively low OSR stems from the fact that, as the OSR increases, the number of ternary taps (i.e., the order of the FIR ternary filter) should be increased to maintain the same frequency response. Hence, a second-order multiple feedback  $\Sigma\Delta M$  is suited to the task of remodulation in the IIR  $\Sigma\Delta M$  filter. Fig. 3 shows a second-order  $\Sigma\Delta M$  used in this filter. This  $\Sigma\Delta M$  has the same structure as the modulator used to encode the impulse response except that it utilizes a single-bit quantizer. The transfer function of the IIR  $\Sigma\Delta$  filter is given below

$$H_{\text{IIR}}(z) = H_{\text{IIRS}}(z) + H_{\text{IIRN}}(z) \quad (4)$$

where  $H_{\text{IIRS}}$  and  $H_{\text{IIRN}}$  are given by

$$H_{\text{IIRS}}(z) = \frac{bz^{-1}}{1 - (1-a)z^{-1}} \quad H_{\text{IIRN}}(z) = \frac{1 - z^{-1}}{1 - (1-a)z^{-1}} \quad (5)$$

Note that  $H_{\text{IIRS}}$  and  $H_{\text{IIRN}}$  represent the signal and noise transfer functions respectively.

The IIR filter coefficient  $a$  was set so as to give the transfer function  $H_{\text{IIRS}}(z)$  a cut-off frequency corresponding to the desired cut-off frequency of the system. The coefficient  $b$  is a gain parameter and should be set so that the overall filtering system has a gain of one. Recall that a ternary filter has a gain factor due to the scaling of the impulse response before modulation. This method of determining the IIR filter coefficients is extremely simple. A more accurate (but more complex) method of obtaining the coefficients of the ternary and the IIR filters can be

found by optimization (in a least square sense) to closely approximate a desired frequency response (see [4] and [5]).

#### IV. PROPOSED FILTER

The final frequency response of the system  $H_{\text{FIL}}$  will be the combination of both the ternary filter and the IIR  $\Sigma\Delta M$  filter. The ternary filter will dominate the low-frequency response, where the IIR  $\Sigma\Delta M$  filter has a relatively flat frequency response. Hence, different frequency characteristics are possible. The IIR  $\Sigma\Delta M$  filter will dominate the higher frequency response, attenuating the coefficient quantization noise of the ternary filter. Using (3) and (4), the system frequency response is given by

$$H_{\text{FIL}}(e^{j\Omega}) = H_{\Sigma\Delta T}(e^{-j\Omega}) \cdot H_{\text{IIR}}(e^{-j\Omega}). \quad (6)$$

From (6) and (4) we have

$$H_{\text{FIL}}(e^{j\Omega}) = H_{\Sigma\Delta T}(e^{-j\Omega}) (H_{\text{IIRS}}(e^{-j\Omega}) + H_{\text{IIRN}}(e^{-j\Omega})) \quad (7)$$

which can explicitly be expressed as follows:

$$H_{\text{FIL}}(e^{j\Omega}) = \frac{G(e^{j\Omega}) [e^{-j\Omega} + e^{-2j\Omega}(b-1)]}{1 - (1-a)e^{-j\Omega}} + \frac{E(e^{j\Omega})}{1 - (1-a)e^{-j\Omega}} \cdot [1 + e^{-j\Omega}(b-3) + e^{-2j\Omega}(3-2b) + e^{-3j\Omega}(b-1)]. \quad (8)$$

The simplicity of the structure and the arithmetic of the single-bit  $\Sigma\Delta M$  filter is its greatest asset. It allows the generation of a filtering system that lends itself to implementation using field-programmable gate Arrays [6]. While the authors have focused on lowpass filtering, the above techniques could be applicable to bandpass filtering. The authors are currently investigating digital bandpass ternary  $\Sigma\Delta M$  filters.

#### V. SIMULATION

A typical speech filter was used as a target for simulating the new filter structure. The filter has a 3-dB cutoff frequency  $f_c$  at 5 kHz and stopband frequency  $f_z$  of 8 kHz. The system was simulated with two different OSRs to highlight the effect that this parameter has on the stopband attenuation. The results are shown in Fig. 4.

The simulated frequency responses of the filter for different OSRs are as expected. Higher OSRs gave higher stopband attenuation, as shown in Fig. 4. In this case, doubling the OSR increased the stopband attenuation by approximately 10 dB. From above, it is evident that increasing the OSR will improve the stopband attenuation. Initial stability tests showed that the system is stable for linear FM and audio signals.

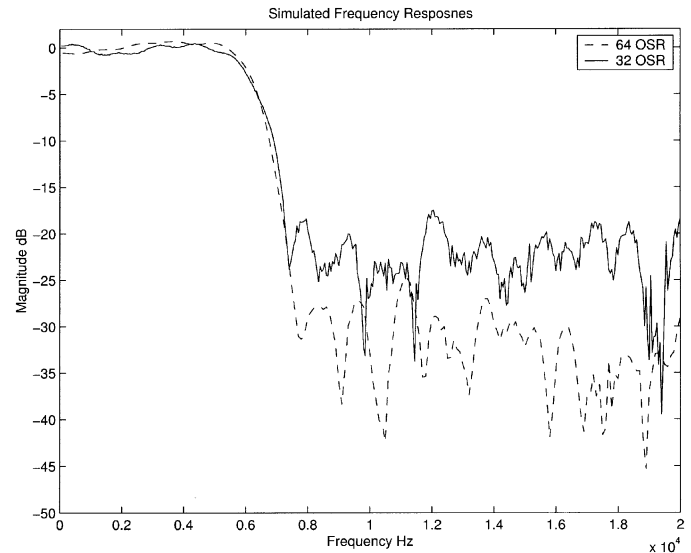


Fig. 4. Block diagram of the simulated signal frequency responses for both 1024 tap (OSR = 32) and 2048 tap (OSR = 64) filters.

#### VI. CONCLUSION

In this letter, a new bit-stream filtering structure is proposed. It consists of a ternary FIR filter cascaded with an IIR  $\Sigma\Delta M$  structure. The simulated frequency response of the overall filter shows that the filter meets the design requirements. Since many of the ternary filter tap values are zero and each nonzero tap requires only very simple multiplication hardware, the system is very resource efficient. Performance enhancement is possible through increasing the oversampling ratio; however, this requires increasing the number of taps and the sampling rate of the system, hence, there is an inherent trade-off between hardware efficiency and performance.

#### REFERENCES

- [1] D. A. Johns and D. M. Lewis, "Design and analysis of delta-sigma based IIR filters," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 233–240, Apr. 1993.
- [2] P. W. Wong, "Fully sigma-delta modulation encoded FIR filters," *IEEE Trans. Signal Processing*, vol. 40, pp. 1605–1610, June 1992.
- [3] S. M. Kershaw, S. Summerfield, M. B. Sandler, and M. Anderson, "Realization and implementation of a sigma-delta bit stream FIR filter," *Proc. Inst. Elect. Eng. Circuits, Devices, Systems*, vol. 143, no. 5, Oct. 1996.
- [4] N. Benvenuto, L. E. Franks, and F. S. Hill, Jr., "Realization of finite impulse response filters using coefficients +1, 0 and -1," *IEEE Trans. Commun.*, vol. COMM-33, Oct. 1985.
- [5] B. Steele and P. O'Shea, "Design of ternary digital filters," in *Proc. 3rd Int. Conf. Information, Communications and Signal Processing*, Oct. 2001.
- [6] J.-L. Cao and P. O'Shea, "A novel FPGA based ternary filter implementation," in *Proc. 3rd Int. Conf. Information, Communications and Signal Processing*, Oct. 2001.
- [7] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma Data Converters: Theory, Design, and Simulation*. Piscataway, NJ: IEEE Press, 1997.