Reconstruction of Oversampled Band-Limited Signals From ΣΔ Encoded Binary Sequences

Søren Hein, *Member, IEEE,* and Avideh Zakhor

***Abstract*—We consider the application of ΣΔ modulators to analog-to-digital conversion. We have previously shown that for constant input signals, optimal nonlinear decoding can achieve large gains in signal-to-noise ratio (SNR) over linear decoding. In this paper we show a similar result for band-limited input signals. The new nonlinear decoding algorithm is based on projections onto convex sets (POCS), and alternates between a time-domain operation and a band limitation to find a signal invariant under both. The time-domain operation results in a quadratic programming problem. The band limitation can be based on singular value decomposition of a certain matrix. We show simulation results for the SNR performance of a POCS-based decoder and a linear decoder for the single loop, double loop and two-stage ΣΔ modulators and for a specific fourth-order interpolative modulator. Depending on the modulator and the oversampling ratio, improvements in SNR of up to 10–20 dB can be achieved.**

I. Introduction

ΣΔ MODULATORS are becoming increasingly important as analog-to-digital converters for relatively low-bandwidth applications such as speech and audio. They are ideal for on-chip VLSI implementation because they require fewer and simpler components than Nyquist-rate converters; this is at the expense of using sampling rates many times faster than the Nyquist rate.

The only nonlinear element in a ΣΔ modulator is a 1-bit quantizer. It is well known that linearizing the quantizer can lead to invalid analytical predictions [1]. Nonetheless, linearization is routinely used in the analysis and design of ΣΔ modulators because the resulting models are simple, and their shortcomings are perceived to be well known. In particular, the decoding of the binary modulator output stream has traditionally been based on linear filtering, and the optimality of this solution has not been questioned. In [2] we point out that linear low-pass decoding is suboptimal because it neglects that the modulator input resulted in the particular observed binary output stream. In [3] we introduce an optimal nonlinear decoding algorithm for constant inputs, and demonstrate large gains over linear decoding.

Manuscript received December 24, 1991; revised April 15, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo. This work was supported by National Science Foundation, PYI grant MIP-9057466, ONR Young Investigator Award N00014-92-J-1732, and Analog Devices.

S. Hein was with the Department of Electrical Engineering and Computer Sciences, University of California-Berkeley, Berkeley, CA 94720. He is now with Siemens AG, Corporate Research, Munich, Germany.

A. Zakhor is with the Department of Electrical Engineering and Computer Sciences, University of California-Berkeley, Berkeley, CA 94720.

IEEE Log Number 9215273.

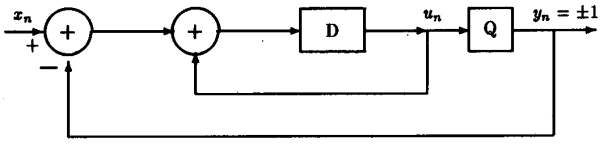

Fig. 1. Discrete-time model of the single loop ΣΔ modulator.

In this paper, we consider nonlinear decoding for band-limited inputs [2], [4]–[6]. We take a signal reconstruction point of view and derive a conceptually feasible, optimal nonlinear decoder applicable to all current ΣΔ architectures. Simulations indicate that our decoder is superior to linear decoding by up to 10–20 dB, depending on the modulator type and the oversampling ratio. These results may serve as upper bounds on attainable performance. Our ultimate goal is a decoder that performs significantly better than a linear decoder, is simple to implement and is as robust towards circuit imperfections as linear decoding.

The paper is organized as follows. Section II presents background information for the decoding problem. Section III describes the reconstruction algorithm. Section IV presents numerical results, and Section V quantifies the effect of circuit imperfections upon decoder performance. Finally, Section VI contains conclusions.

II. Background

In this paper we consider four representative ΣΔ modulator architectures. The simplest one is known as the single loop modulator and is shown in Fig. 1. It consists of an integrator embedded in a nonlinear feedback loop, which also includes a one-bit quantizer Q. Q is defined as the signum function of its argument, and can be viewed as adding a noise sequence to its input sequence. The noise sequence is often assumed to be approximately white and uncorrelated with the input sequence $\{x_n\}$, although this assumption is inaccurate.

Fig. 2 shows Candy's double loop ΣΔ modulator, which contains two integrators [7]. Similarly, Fig. 3 shows the two-stage MASH modulator which also contains two integrators, but in a cascaded configuration involving two quantizers. Finally, Fig. 4 shows a generic interpolative ΣΔ modulator that includes a discrete-time open-loop filter $H(z)$. The single loop modulator is an interpolative modulator with $H(z) = z^{-1}/(1-z^{-1})$. We will also consider the specific fourth-order modulator defined in [8] by

$$H(z) = z^{-1} \cdot \sum_{i=0}^{4} N_i z^{-i} \Big/ \sum_{i=0}^{4} D_i z^{-i} \qquad (1)$$

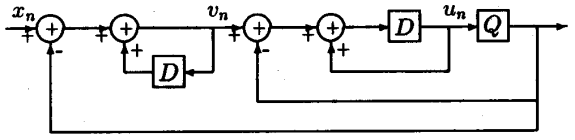

Fig. 2. Discrete-time model of the double-loop ΣΔ encoder.

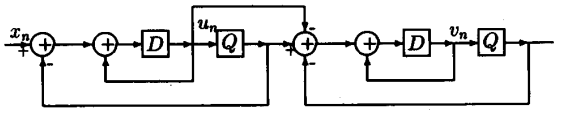

Fig. 3. Discrete-time model of the two-stage ΣΔ encoder.

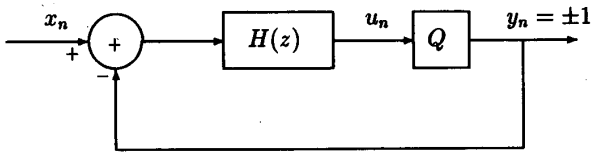

Fig. 4. Discrete-time model of the interpolative ΣΔ encoder.

where $N_0 = 0.8653$, $N_1 = -2.2692$, $N_2 = 2.0064$, $N_3 = -0.59714$, $N_4 = 0.000035$, $D_0 = 1$, $D_1 = -3.99646$, $D_2 = 5.992922$, $D_3 = -3.996460866$, $D_4 = 1.000000433$. $H(z)$ is in general chosen to be a low-pass filter whose pass-band corresponds to the frequency range in which the input signal is concentrated.

To obtain a digital approximation to the analog modulator input, the digital output is traditionally lowpass filtered and decimated. The quality of the approximation depends on the oversampling ratio (OSR), which is defined as the ratio between the sampling rate and the Nyquist rate of the input. The single loop, double loop and two-stage modulators can operate at arbitrary OSR's, whereas the fourth-order modulator (1) is designed specifically for OSR = 48.

III. Reconstruction Algorithm

In this section we describe our proposed reconstruction algorithm in the context of the general interpolative ΣΔ encoder shown in Fig. 4. Although the double-loop and two-stage encoders are not interpolative, the algorithm easily extends to these architectures. Section III-A describes the POCS algorithm [9], [10] which iteratively enforces two constraints on an input signal estimate: First, that the input signal resulted in the observed output signal, and second, that it is band-limited. Section III-B shows that finding an input sequence that generates the observed output signal can be treated as a quadratic programming (QP) problem. Section III-C reviews our singular value decomposition (SVD) based approach to band limitation. Section III-D describes modifications to the POCS algorithm for sliding-block decoding. Section III-E compares our proposed algorithm to that of Thao and Vetterli [11]–[14].

A. Projections onto Convex Sets

In this section we briefly review the POCS algorithm as it applies to our decoding problem. For any binary output signal $\mathbf{y} = \{y_0, \cdots, y_{N-1}\}$ of an interpolative ΣΔ encoder, we define the set S_1 to contain all input signals $\mathbf{x} = \{x_0, \cdots, x_{N-1}\}$ that result in $\mathbf{y}$ when applied to the encoder. We also define the set S_2 to contain all N-sample signals $\mathbf{x}$ that are band-limited; the precise meaning of this is discussed in Section III-C. To estimate the input signal optimally, we must find a signal $\hat{\mathbf{x}} \in S_1 \cap S_2$. This problem formulation makes the POCS algorithm a natural choice [9]. We show below that S_1 and S_2 are convex[1] as assumed by POCS. If we denote the orthogonal projections onto S_1 and S_2 by P_1 and P_2, respectively, then the theory of POCS states that an element $\hat{\mathbf{x}} \in S_1 \cap S_2$ can be found from any initial guess $\mathbf{x}_0$ by the iteration [9]

$$\mathbf{x}_{n+1} = (P_2 \circ P_1)\mathbf{x}_n, n \geq 0, \hat{\mathbf{x}} = P_1\hat{\mathbf{x}} = P_2\hat{\mathbf{x}} = \lim_{n\to\infty} \mathbf{x}_n.$$

In the following sections we discuss the projections separately.

B. Time-Domain Projection

In this section we consider the time-domain projection P_1 of a signal $\mathbf{x}$ onto the space S_1 of sequences that generate a given output sequence $\mathbf{y}$. We derive a time-domain characterization of the ΣΔ encoder, and show that the projection P_1 results in a QP problem.

We denote the impulse response of $H(z)$ by $\{h_0, h_1, \cdots\}$; h_0 must be zero to avoid delay-free loops. We define the $N \times N$ lower-triangular Toeplitz matrix

$$\mathbf{H} = \begin{bmatrix} h_1 & 0 & \cdots & 0 & 0 \\ h_2 & h_1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ h_{N-1} & h_{N-2} & \cdots & h_1 & 0 \\ h_N & h_{N-1} & \cdots & h_2 & h_1 \end{bmatrix}.$$

For an Mth order encoder, we define an $(M+1)$-dimensional state vector $\mathbf{s}$[2], which we assume to be known in this section. We also define an $N \times (M+1)$ zero-input response matrix $\mathbf{Z}$ so that if the filter $H(z)$ is driven open-loop with initial state $\mathbf{s}$ and zero input over N samples, the filter output is $\mathbf{Zs}$. We can then write the quantizer input vector $\mathbf{u} = \{u_0, u_1, \cdots, u_{N-1}\}$ as

$$\mathbf{u} = \mathbf{H}(\mathbf{x} - \mathbf{y}) + \mathbf{Zs}.$$

Our only knowledge of any particular quantizer input u_n is in the form of its sign, $Q(u_n)$. This knowledge provides us with a series of bounds on linear combinations of sample values $\{x_n\}$. To cast this in matrix notation, we define an $N \times N$ diagonal matrix with ± 1 entries on its diagonal, $\mathbf{Q} = -\text{diag}(y_1, y_2, \cdots, y_N)$. The bounds imposed by the quantizer outputs are then

$$\mathbf{QHx} \leq \mathbf{Q}(\mathbf{Hy} - \mathbf{Zs}) \tag{2}$$

where the inequality sign is to be taken coordinate-wise. Equation (2) establishes a time-domain description of the

[1] A set S is convex if for all $\mathbf{a}, \mathbf{b} \in S$, $\alpha\mathbf{a} + (1-\alpha)\mathbf{b} \in S$ for any $0 < \alpha < 1$.

[2] The overall delay in $H(z)$ in general adds one to the system order. For the single-loop encoder, however, the filter delay and the overall delay can be combined into one.

interpolative encoder.[3] We can show that the set S_1 of signals satisfying (2) is convex. Specifically, if we consider two sample vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ that both satisfy (2), then for any $0 < \alpha < 1$

$$\mathbf{Q}\mathbf{H}[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2] \leq \alpha\mathbf{Q}(\mathbf{H}\mathbf{y} - \mathbf{Z}\mathbf{s}) + (1-\alpha)\mathbf{Q}(\mathbf{H}\mathbf{y} - \mathbf{Z}\mathbf{s}) = \mathbf{Q}(\mathbf{H}\mathbf{y} - \mathbf{Z}\mathbf{s}).$$

We will adopt the 2-norm as our performance and projection metric, that is, we define the Signal-to-Noise Ratio[4] (SNR) to be $10\log_{10}(E_{\text{signal}}/E_{\text{noise}})$ where

$$E_{\text{signal}} = \sum_{n=0}^{N-1} x_n^2, \qquad E_{\text{noise}} = \sum_{n=0}^{N-1} (x_n - \hat{x}_n)^2. \quad (3)$$

Projecting onto S_1 in the 2-norm is equivalent to finding the signal $\hat{\mathbf{x}}$ that satisfies (2) and minimizes the distance $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$. This criterion and equation (2) together constitute a linearly constrained QP problem, which can be solved with techniques available in the literature [15].

In practice, sample sizes N on the order of thousands are desirable to increase the accuracy of the band limitation projection described in Section III-C below. Because of the large computational complexity of the QP problem, we will derive an approximation to P_1, which takes advantage of the fact that $\mathbf{H}$ is lower-triangular. The idea is to solve a number of L-dimensional QP problems ($L < N$) rather than a single N-dimensional one, by dividing the signal $\mathbf{x}$ into L-sample QP subblocks. In the simplest set-up, the subblocks are not overlapping. The small QP problems are solved in chronological order, and no subblock is allowed to change the estimates of previous subblocks. This approach is illustrated in Fig. 5(a) and is described mathematically in the appendix. The problem with this set-up is that large changes tend to be necessary at the beginning of each subblock because $\mathbf{H}$ is lower-triangular, and so each bound violation must be corrected using only samples before the violation. Bound violations at the beginning of a subblock therefore require larger sample modifications than do violations towards the end of a subblock. In our preferred set-up the subblocks are partially overlapping, and thus the optimization of each subblock takes into account a portion of the subblock immediately following it. This approach is shown in Fig. 5(b) and is also described mathematically in the appendix. The overall QP estimate of an N-sample block is obtained by concatenating parts of the estimates resulting from each of the QP subblocks. Specifically, we use the estimate on that part of each subblock that does not overlap with the immediately following subblock. The overlapping portion of each subblock with the next subblock is used to improve upon the initial estimate of the QP solution for the following subblock.

[3] Similar descriptions can be found for noninterpolative encoders. The main differences for such encoders are as follows: For the double-loop encoder, the impulse response matrix seen by the input $\mathbf{x}$ and the output $\mathbf{y}$ are different. For the two-stage encoder, we assume as in [3] that we have access to both quantizer outputs; each quantizer gives a series of bounds, and we can enforce the two sets of bounds separately. For predictive coders such as the Delta encoder, the impulse response matrix is the identity matrix, and as a result the time-domain projection becomes trivial.

[4] This ratio is sometimes also referred to as the signal-to-noise plus distortion ratio (SDR).

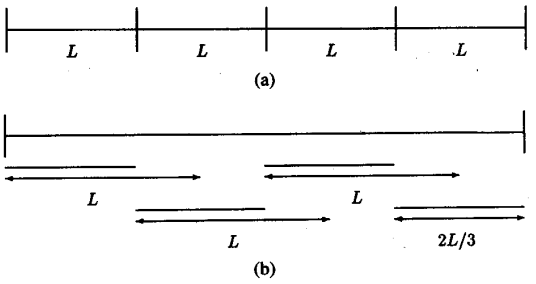

Fig. 5. Two ways of dividing an N-dimensional QP problem into smaller L-dimensional problems. (a) Nonoverlapping blocks. (b) Blocks overlapping by $\frac{1}{3}$; the marked parts of each block are used in the final solution estimate.

The choice of QP block size L is a trade-off between the speed with which a solution is obtained, and the closeness of the solution to the true 2-norm solution. The closeness to the 2-norm is important because the frequency-domain projection will be done in the 2-norm, and the POCS algorithm only guarantees convergence if the projections are done in the same norm.

The computational complexity of the time-domain projection depends on the number of samples N and the QP block size L. To quantify the complexity, we prefer empirical run-time observations over analytical bounds, as the bounds are worst-case bounds, and the particular problem we are addressing has significant structure that the bounds do not take into account. Empirically, we find each QP subblock problem to depend cubically on L, so the overall empirical complexity of the time-domain projection is $O(\frac{N}{L} \cdot L^3) = O(NL^2)$.

C. Frequency-Domain Projection

In this section we consider the frequency-domain projection P_2 of a signal $\mathbf{x} = \{x_0, \cdots, x_{N-1}\}$ onto the set S_2 of signals that are band-limited in a sense to be defined. This projection raises some general questions about band limitation that are discussed in [16] and are summarized below.

The two standard techniques for band limitation are the discrete Fourier transform (DFT) and linear filtering. However, the DFT is only accurate when the signal frequencies are bin frequencies for the DFT, which cannot in general be assumed. We showed in [16] that the frequency of a sinusoidal baseband signal can be chosen half way between DFT bin frequencies such that the SNR is as low as

$$\text{SNR}_{\min} \approx 10\log_{10}\left(\frac{N\pi^2}{4\text{OSR}}\right).$$

For $N = 4096$ and $\text{OSR} = 48$, we get $\text{SNR}_{\min} = 23$ dB, and this minimum SNR only increases by 3 dB/octave with the sample size. Windowing can be used to suppress spectral leakage in the DFT, but this has the side-effect of smearing the signal spectrum significantly as shown in [16].

Linear filtering for moderate-length block-oriented processing has the disadvantage that we must pad the block of samples with zeros to obtain a sequence on which a linear filter can operate, and the filtered result exhibits large edge effects due to this padding. In practice, linear filtering is often performed continuously on a stream of modulator outputs; however, for

the block-oriented processing that we consider for POCS, we only have access to a finite output block of moderate length. In addition, similarly to any practical band limiter, a linear filter can only approach the ideal low-pass transfer function that has unity baseband gain and zero high-frequency gain, and the deviations from ideality result in errors. Finally, linear filtering is not a projection unless the filter transfer function only takes on the values 1 and 0.

Our preferred, general technique for band limitation is described in [16] and is related to known algorithms for band-limited extrapolation [17]. In [16] we also discuss the fact that any finite set of samples can be viewed as samples of infinitely many different infinite-extent sequences with band-limited discrete time Fourier transforms (DTFT's), and thus it is not clear how to define band limitation for general finite sequences. This conceptual problem is resolved by using energy and dimensional considerations. Our band-limitation technique is based on the SVD of an $N \times N$ symmetric Toeplitz matrix $\mathbf{L}$ of samples of the impulse response of an ideal low-pass filter, $\ell_{mn} = \sin[\sigma(m-n)]/[\pi(m-n)]$ where $\sigma = 1/\text{OSR}$. The SVD leads to the truncated discrete prolate spheroidal sequences (DPSS's) of N-dimensional vectors $\{\mathbf{u}_0, \cdots, \mathbf{u}_{N-1}\}$, which have been studied extensively by Slepian [18] and others. The truncated DPSS's form an orthonormal basis of $\mathcal{R}^N$. Approximately $r = N/\text{OSR}$ of the singular values of $\mathbf{L}$ are close to 1, while the remaining ones are close to 0. The truncated DPSS's for singular values close to 1 and 0 are analogous to the baseband and high-frequency (HF) complex exponential basis functions, respectively, of the DFT. Our SVD-based algorithm projects signals onto the space S_2 spanned by the "baseband" DPSS's

$$\hat{\mathbf{x}} = \sum_{n=1}^{r} (\mathbf{x}^T \mathbf{u}_n) \cdot \mathbf{u}_n, r \approx N/\text{OSR}. \tag{4}$$

The projection (4) is in the sense of the 2-norm, and the space S_2 is linear and thus convex. We consider the signals in the range space of $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ to be band-limited. The SVD required to obtain the vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ only needs to be done once, and can be implemented efficiently using the Lanczos algorithm [19] and exploiting the Toeplitz form of $\mathbf{L}$. The computational complexity of the projection (4) is $O(N^2/\text{OSR})$, which is found to be comparable in speed to a fast Fourier transform (FFT) band-limitation for a sample size of 4K (1K = 1024).

D. Sliding Block Decoding

The POCS algorithm described in Sections III-A–III-C is suitable for decoding blocks of N samples when the initial encoder state is known. However, we find empirically that the algorithm suffers from larger estimation inaccuracy towards the edges of an N-sample block than on its middle. To reduce edge effects, we therefore also consider a sliding block set-up where the N-sample blocks overlap by 50% as shown in Fig. 6. Each block is decoded using the iterative POCS algorithm, and upon convergence, the initial state of the following block is estimated. The state estimation is done by simulating the behavior of an encoder that starts in the initial state of the current block and has the estimated input sequence as its input. The overall input estimate is the concatenation of the input estimates on the middle halves of the N-sample blocks. The SNR is calculated in the time domain over the middle half of each N-sample block, and the overall SNR is defined as the average of the SNR's on the contributing blocks.

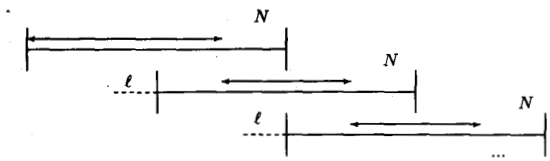

Fig. 6. 50% overlap of the N-sample blocks on which the POCS iterations operate. The marked parts of each block are used in the overall solution. The dotted lines show an extension of the QP projection described in the text.

The accuracy of the time-domain and frequency-domain projections can be improved upon in two ways by extending the idea of overlapping blocks. First, for all but the first N-sample block, the QP operation can be started a number ℓ samples before the beginning of the N-sample block and extend over $N+\ell$ samples. This implies that the QP operation on the previous N-sample block includes the first ℓ samples of the present N-sample block, and reduces the effects of early bound violations for the same reason that overlapping the QP subblocks reduces these effects. Second, the SVD projection tends to be most accurate on the middle of an N-sample block, so results from the part of the previous N-sample block that overlaps with the current N-sample block can be used to improve upon the accuracy of the initial estimate for the current block. On samples where no better estimates are available from previous N-sample blocks, the initial estimate is chosen to be the binary output sequence of the encoder.

E. Comparison to Thao and Vetterli's Algorithm

In this section we compare our proposed algorithm [2], [4]–[6], to the algorithm proposed by Thao and Vetterli [11]–[14], which was developed independently. Both algorithms are POCS-based. Sections III-E-1 and III-E-2 discuss the time-domain and frequency-domain projections, and Section III-E-3 discusses the modes of operation for the two algorithms.

1) Time-Domain Projection: Thao and Vetterli develop an approximate method for performing the time-domain projection, which is valid for certain nth order multistage $\Sigma\Delta$ encoders as well as for other encoders. The idea is to only enforce a subset of the bounds imposed by the encoder outputs, and to recursively attempt to guess which bounds are active [15]. For orders n above 1, the authors' algorithm only yields an approximate projection onto the convex subspace spanned by the particular bounds that are assumed active; the algorithm is thus approximate both in the way that the active constraints are chosen, and in the way that the resulting optimization problem is solved. The advantage of the authors' interesting approach over the more accurate approach of performing the full projection is a decrease in computational complexity. The authors do not comment on the achieved gain over conventional, efficiently implemented QP, or an approximate

QP approach such as the one described in Section III-B, which uses all available bounds. Nor do they describe the loss of accuracy resulting from their approximations. The authors do not quantify the fraction of bounds that are typically used by their algorithm, but their most general algorithm described in [12] requires that the active bounds be spaced apart by at least n samples. The authors do not state whether their algorithm permits a range of trade-offs between computational speed and accuracy. The authors' algorithm only works for a subset of $\Sigma\Delta$ encoders, most notably multi-stage encoders whose individual stages all have their open-loop filter poles at dc. The approach is thus not applicable to the general class of interpolative encoders with filter poles over signal baseband [8] such as the fourth-order encoder considered in Section IV. Because of the assumption of dc filter poles, it would be interesting to see simulation results for the sensitivity of the authors' algorithm towards unknown nonidealities in the encoder filters such as integrator leak.

2) Frequency-Domain Projection: In order to have access to an ideal low-pass filtering for the frequency-domain projection, Thao and Vetterli assume that the input signal is periodic. They also assume that the period is equal to the number of samples N to which the decoder has access [12]. In effect, they thus assume that the input frequency is known.

3) Mode of Operation: With regard to a continuous mode of operation, Thao and Vetterli assume that the integrators states are reset to zero at the beginning of each conversion cycle. While our algorithm can also work in this mode, we found that increased accuracy is obtained by overlapping the blocks and propagating estimates of the encoder states, as described in Section III-D; this enables us to only count estimation errors on the middle half of each POCS block. Another potential advantage of not periodically resetting the integrator states is that encoder transients are avoided.

IV. Results

This section contains numerical results for single block and sliding block decoding for a number of $\Sigma\Delta$ encoders, including the single-loop, double-loop and two-stage encoders and the specific fourth-order encoder given by equation (1) [8]. For single block simulations, the SNR is calculated over the middle half of an N-sample block, and for sliding block simulations, the SNR is calculated as explained in Section III-D. We point out that our SNR measure includes both low frequency and high frequency signal errors.

Throughout the section, the input signal is a sinusoid whose amplitude is varied and whose frequency is $\alpha = 9/32$ times the largest signal frequency for which the modulator is designed to operate, that is, $9/32$ of half the Nyquist rate.[5] This frequency is a bin frequency for the DFT to enable comparisons between the POCS decoders whose band limitations are based on the DFT and on the SVD (4), respectively. When we use a bin frequency, the DFT will perform better than the SVD, and so the shown DFT-based experimental results will be superior to the corresponding SVD-based ones. However, the SVD performs significantly better than the DFT on non-bin frequencies, and signal frequencies cannot in general be assumed to be bin frequencies.

[5]Other simulations show that similar results are obtained for other signal frequencies. For brevity, one representative frequency is chosen.

TABLE I
Summary of Peak SNR Performance Results for the Single-Loop, Double-Loop, Two-Stage and Fourth-Order Encoders; the Encoder Components are Assumed Ideal. Both the POCS and the Linear Decoders can be Based on Either the DFT or the SVD. The Linear Decoders are Generally Limited by Signal Frequency Errors

			Peak SNR (dB)	
Encoder	Decoder	OSR	DFT	SVD
Single-loop	POCS	64	61	56
		128	84	67
	Linear	64	37	37
		128	44	44
Double-loop	POCS	64	79	65
		128	92	83
	Linear	64	37	37
		128	44	44
Two-stage	POCS	64	86	85
		128	101	94
	Linear	64	31	31
		128	38	38
Fourth-order	POCS	48	93	83
	Linear	48	84	64

For the shown results, the sample size for each block is 4K. For sliding block simulations, we consider three 4K blocks that overlap by 50% for a total of 8K samples. The overlap between QP subblocks is $L/3$ where $L = 192$ for the fourth-order encoder and $L = 96$ for the single-loop, double-loop and two-stage encoders. The initial state vector $\mathbf{s}$ for the first block is assumed to be known and is set to zero. The number of singular vectors r in (4) is 94, 74 and 39 for oversampling ratios OSR = 48, 64 and 128, respectively. The number of POCS iterations is 12.

Sections IV-A–IV-D contain results for the single-loop, double-loop, two-stage and fourth-order encoders, respectively. The main results are summarized in Table I; we make comments on the dependence of the SNR upon the OSR in the individual sections below. In Section IV-E we compare our SNR metric and performance results to existing results in the literature.

A. Single-Loop Modulator

We first consider single block decoding with the SVD-based band limitation (4) as the frequency domain projection, for the two oversampling ratios 64 and 128. Fig. 7 shows SNR curves for the SVD-based POCS decoder, as well as for a low-pass decoder that uses (4). The figure shows that particularly for large input amplitudes, the POCS algorithm is clearly superior to the linear decoder. For OSR = 64, the peak SNR's are 56 dB and 37 dB, and for OSR = 128, the peak SNR's are 67 dB and 44 dB for the SVD-POCS and SVD-alone decoders, respectively[6].

[6]We have chosen the peak SNR as a convenient measure of performance. For certain applications, other measures such as the dynamic range are often used to complement or replace the peak SNR measure.

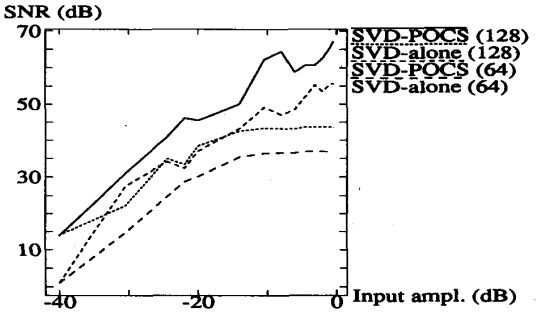

Fig. 7. SNR as a function of input amplitude for the single-loop encoder with single block decoding. The two decoders are an SVD-based low-pass decoder and the SVD-based POCS decoder.

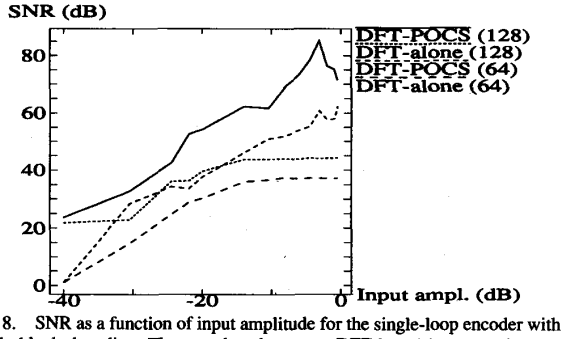

Fig. 8. SNR as a function of input amplitude for the single-loop encoder with single block decoding. The two decoders are a DFT-based low-pass decoder and the DFT-based POCS decoder.

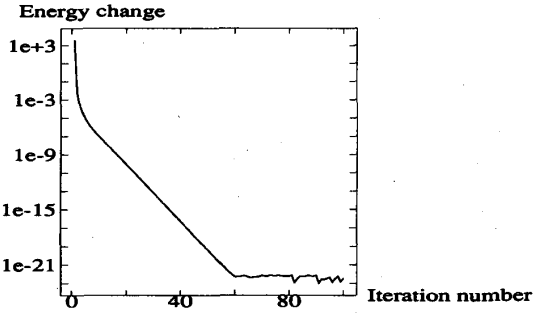

Fig. 9. Energy modification performed by the frequency-domain projection as a function of POCS iteration number. The decoder is the DFT-POCS one.

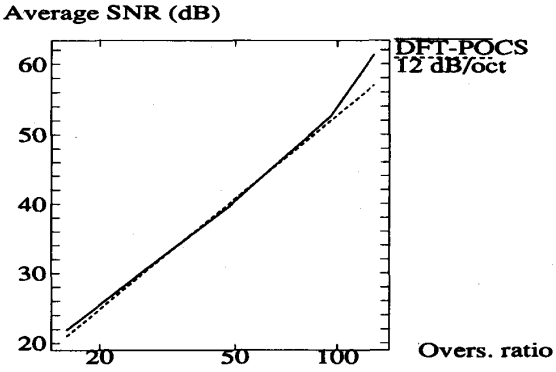

Fig. 10. Average SNR as a function of oversampling ratio for the single-loop encoder with single-block decoding and the DFT-POCS decoder. Also shown is a line with 12 dB/octave slope.

We have carried out a set of simulations under an identical set-up to that of Fig. 7, except that we consider a sliding-block decoder. We find that the SNR curves for the sliding-block decoder are within 1–2 dB of the corresponding curves in Fig. 7, indicating that state uncertainty implies little loss in SNR, so the same conclusions hold as for Fig. 7.

For comparison, Fig. 8 shows SNR results for the POCS algorithm when using the DFT rather than the SVD-based method (4) as the band limitation. The figure is valid for single block decoding, but another set of simulations for sliding-block decoding shows that the difference between corresponding single block and sliding-block curves is 0.1–0.2 dB. For OSR = 64, the peak SNR's are 61 dB and 37 dB, and for OSR = 128, the peak SNR's are 84 dB and 44 dB for the DFT-POCS and DFT-alone decoders, respectively.

Comparing Figs. 7 and 8, the SVD-based POCS method is up to 20 dB superior to the DFT-alone and SVD-alone methods. We see that the SVD-based POCS method mostly loses 2–4 dB compared to the DFT-based POCS method for OSR = 64 and up to about 10 dB for OSR = 128. We find that the linear decoders are all limited by signal frequency errors.

To demonstrate the convergence of our algorithm, Fig. 9 shows the energy modification performed by the frequency-domain projection as a function of POCS iteration number. The curve is valid for single-block decoding with the DFT-POCS algorithm and an input amplitude of 0.10. We see that the algorithm finds an estimated input within about 10^{-23} of satisfying both the time-domain and frequency-domain constraints; the small error may be attributed to machine inaccuracy in the DFT.

To assess the SNR dependence of the DFT-POCS algorithm upon the OSR, we have obtained simulation results for OSR's of 16, 32, 48, 64, 96 and 128. To combine the SNR performance over the range of input amplitudes into one indicator, we average the SNR over 14 fixed amplitudes between 0.01 and 0.95.[7] This average SNR should of course only be taken as a relative rather than absolute performance indicator, but it does permit comparison across OSR's. Fig. 10 shows the trade-off between the average SNR and the OSR. We see that the slope of the curve is approximately 12 dB/octave; this result is further commented upon in Section IV-E-4.

B. Double-Loop Modulator

For the double-loop encoder, we consider single block decoding with the SVD-based band limitation (4) as the frequency domain projection, for the two oversampling ratios 64 and 128. For the SVD-based POCS decoder, we find that particularly for large input amplitudes, the POCS algorithm is clearly superior to the linear decoder. The peak SNR's are indicated in Table I. Fig. 11 shows SNR results for the POCS algorithm when using the DFT rather than the SVD-based

[7] The amplitudes are the same ones used to generate the plots in this section, specifically, 0.01, 0.03, 0.06, 0.08, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95.

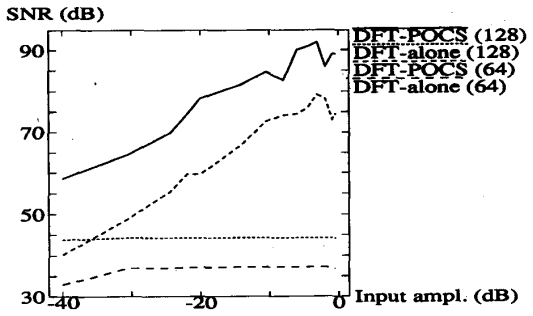

Fig. 11. SNR as a function of input amplitude for the double-loop encoder with single block decoding. The two decoders are a DFT-based low-pass decoder and the DFT-based POCS decoder.

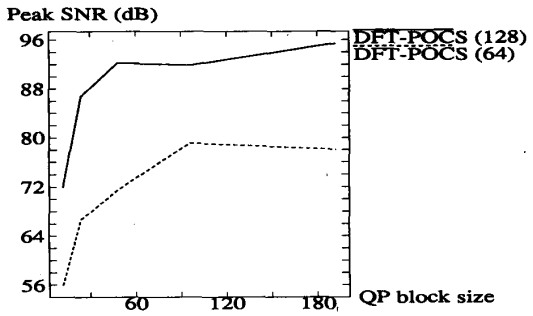

Fig. 12. Peak SNR as a function of QP block size for the double-loop encoder with single-block DFT-POCS decoding.

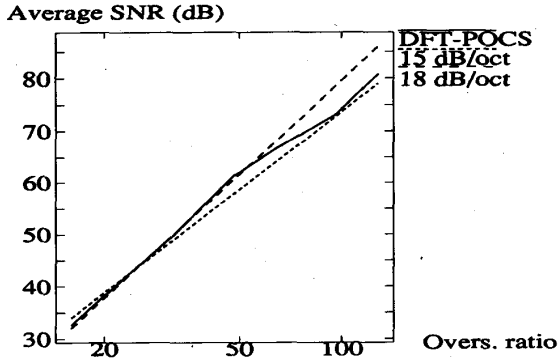

Fig. 13. Average SNR as a function of oversampling ratio for the double-loop encoder with single-block decoding and the DFT-POCS decoder. Also shown are lines with 15 dB/octave and 18 dB/octave slopes.

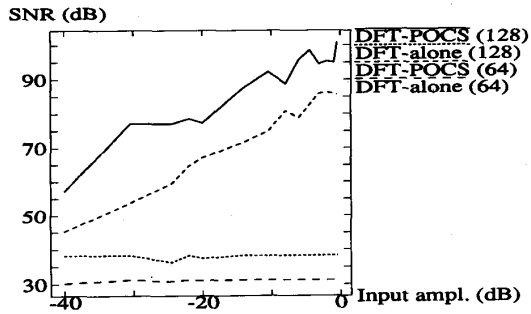

Fig. 14. SNR as a function of input amplitude for the two-stage encoder with single block decoding. The two decoders are a DFT-based low-pass decoder and the DFT-based POCS decoder.

method (4) as the band limitation. The figure is valid for single block decoding, but another set of simulations for sliding-block decoding shows that the difference between corresponding single block and sliding-block curves is about 3 dB. For OSR = 64, the peak SNR's are 79 dB and 37 dB, and for OSR = 128, the peak SNR's are 92 dB and 44 dB for the DFT-POCS and DFT-alone decoders, respectively.

In comparing results for the SVD-based and DFT-based algorithms, we find that the SVD-based POCS method is up to 40 dB superior to the DFT-alone and SVD-alone methods. We see that the SVD-based POCS method mostly loses 5–10 dB compared to the DFT-based POCS method for OSR = 64 and up to 10–15 dB for OSR = 128. We find that the linear decoders are all limited by signal frequency errors.

As an example of the influence of the QP block size L, Fig. 12 shows the peak SNR of the DFT-POCS decoder as a function of L. This figure justifies our choice of $L = 96$ for the double-loop encoder.

Fig. 13 shows the trade-off between the average SNR and the OSR, obtained as in Section IV-A. We see that the slope of the curve is approximately 18 dB/octave on most parts of the curve, but the curve suffers a dip in average SNR around an OSR of 80, and the overall slope thus appears to be slightly less than 18 dB/octave. This result is further commented upon in Section IV-E-4.

C. Two-Stage Modulator

For the two-stage encoder, we consider single block decoding with the SVD-based band limitation (4) as the frequency domain projection, for the two oversampling ratios 64 and 128. For the SVD-based POCS decoder, we find that particularly for large input amplitudes, the POCS algorithm is clearly superior to the linear decoder. The peak SNR's are indicated in Table I. For comparison, Fig. 14 shows SNR results for the POCS algorithm when using the DFT rather than the SVD-based method (4) as the band limitation. For OSR = 64, the peak SNR's are 86 dB and 31 dB, and for OSR = 128, the peak SNR's are 101 dB and 38 dB for the DFT-POCS and DFT-alone decoders, respectively.

In comparing results for the SVD-based and DFT-based algorithms, we find that the SVD-based POCS method is up to 40 dB superior to the DFT-alone and SVD-alone methods. We see that the SVD-based POCS method mostly loses 5–10 dB compared to the DFT-based POCS method for OSR = 64 and OSR = 128. We find that the linear decoders are all limited by signal frequency errors.

Fig. 15 shows the trade-off between the average SNR and the OSR, obtained as in Section IV-A. We see that the slope of the curve is approximately 18 dB/octave up to an OSR of about 80, and then appears to drop slightly. This result is further commented upon in Section IV-E-4.

D. Fourth-Order Modulator

We first consider single block decoding with the SVD-based band limitation (4) as the frequency domain projection. Fig. 16

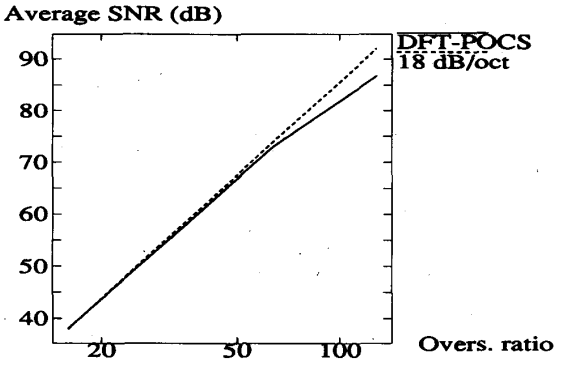

Fig. 15. Average SNR as a function of oversampling ratio for the two-stage encoder with single-block decoding and the DFT-POCS decoder. Lines with 15 and 18 dB/octave slopes are also shown.

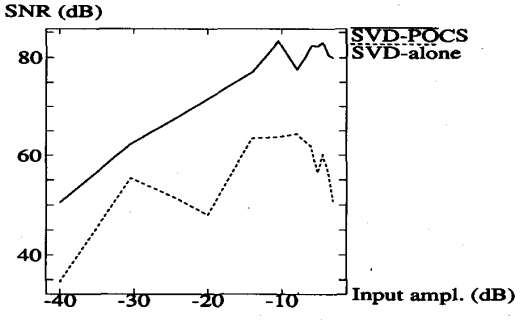

Fig. 16. SNR as a function of input amplitude for a specific fourth-order interpolative encoder. The two decoders are the SVD-based low-pass decoder and the SVD-based POCS decoder.

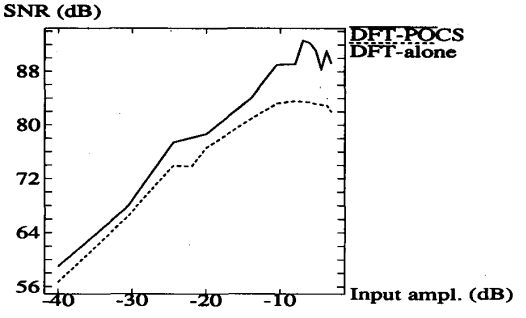

Fig. 17. SNR as a function of input amplitude for a specific fourth-order interpolative encoder. The two decoders are the DFT-based low-pass decoder and the DFT-based POCS decoder.

shows SNR curves for the SVD-based POCS decoder, as well as for a low-pass decoder that only uses (4). The figure shows that the peak SNR is 83 dB for the SVD-POCS decoder and 64 dB for the SVD-alone decoder. The gain of the SVD-POCS decoder is 10–20 dB.

For comparison, Fig. 17 shows SNR results for single block decoding for the POCS algorithm when using the DFT rather than the SVD-based method (4) for band limitation. The peak SNR's are 93 dB for the DFT-POCS decoder and 84 dB for the DFT-alone decoder, and the SNR gain of the DFT-POCS decoder is 5–10 dB.

Other simulation results show that under the sliding block set-up, both the SVD-POCS and the DFT-POCS decoders lose 1–2 dB compared to the corresponding single block decoders. This means that state uncertainty results in little SNR loss.

The DFT-alone curve in Fig. 17 is generally 15–30 dB above the SVD-alone curve in Fig. 16. As explained above, we would expect the DFT-alone method to be superior because we are using a bin frequency signal as the input. However, the difference between the DFT-alone and SVD-alone curves could have been made smaller by a different choice of the number of singular vectors for the SVD band limitation (4). The number of singular vectors was chosen to optimize the SVD-POCS curves rather than the SVD-alone ones.

E. Comparison with Existing Results

In this section we compare our results to existing results in the literature. In Section IV-E-1 we discuss the choice of SNR definition, and in Section IV-E-2 we present simulation results to illustrate the performance of the POCS algorithm under different SNR definitions. In Section IV-E-3 we discuss the possibility that the POCS algorithm may be capable of suppressing the spectral tones that plague linear decoders for low-order encoders. In Section IV-E-4 we compare our results to those of Thao and Vetterli.

1) SNR Definition: To enable comparisons with existing results, we change our definition of SNR from the current definition (3), which counts the 2-norm difference between the sinusoidal input signal and its decoded version as noise. Our argument for using (3) is that a decoder should ultimately be judged on the total amount of noise that it leaves on its input estimate—otherwise, any additional circuitry to further reduce the noise should be viewed as part of the decoder. Nonetheless, the changes that we consider are:

1) Exclusion of the noise contribution from so-called linear errors in the SNR calculation. Linear errors are the errors at the frequency of the sinusoidal input signal. This modification is in agreement with [8], [20], [21].
2) Using an N-point Hanning window on the difference signal between the input and the decoder output [22] before computing the noise power. This modification is in agreement with [8].

Exclusion of linear errors is generally justified by noting that they can be compensated for with simple scaling. However, there are two reasons that this justification is not valid in a $\Sigma\Delta$ context:

1) The required scaling may be frequency-dependent. Signal distortion results from the nonunity signal transfer function of the encoder, and from a linearized point of view, the encoder has a signal transfer function between its input and output of $H_X(z) = H(z)/[1 + H(z)]$, where $H(z)$ is the open-loop transfer function. In general the signal transfer function is not constant over baseband, and thus a simple frequency-independent constant gain is insufficient to equalize it. On the other hand, an equalization filter can be designed to compensate for the ripples in $H_X(z)$ over baseband, such that the

product of its transfer function with $H_X(z)$ approaches unity over baseband.

2) The signal frequency error in general depends nonlinearly on the signal amplitude and phase as well as on the presence of other signals. This is because a $\Sigma\Delta$ encoder is a nonlinear circuit.

To illustrate the importance of signal frequency error, we consider a situation where the signal frequency component at the encoder output equals a constant K times a sinusoidal input signal. We assume for simplicity that K is real. The signal frequency error energy equals $(1-K)^2$, so if the SNR is defined to include signal frequency error, the SNR is limited to $-20\log_{10}|1-K|$. For instance, if $K = 1+10^{-4}$, the SNR cannot exceed 80 dB. For a linear filter, an inaccuracy of $1-K = 10^{-4}$ corresponds to a ripple of 0.0009 dB. This example shows that the passband demands on the equalizing filter are quite stringent. In addition, the filter must act as a good low-pass filter outside of baseband. Even if these requirements are met, the nonlinear signal distortion inherent in the encoder cannot be cancelled.

Although our POCS decoding algorithm does not explicitly take signal frequency errors into account, we have found numerically that the time-domain constraint acts to significantly reduce these errors. The POCS algorithm thus avoids the need to explicitly design a high-precision linear filter, and furthermore the time-domain constraint implies that the algorithm does not neglect the nonlinear signal frequency errors. These observations are further verified in the following section.

2) Simulation Results: Fig. 18 is analogous to Fig. 8 and shows SNR results for the single-loop encoder when the SNR is obtained by discarding linear errors. The band limitation is DFT-based. The DFT-alone curves can be taken as upper bounds on the performance of linear filtering. For OSR = 64, the peak SNR's are 63 dB for the DFT-POCS decoder and 58 dB for the DFT-alone decoder. For OSR = 128, the peak SNR's are 84 dB and 66 dB, respectively. Fig. 18 shows that when excluding linear errors, the POCS method gains about 3–7 dB over the DFT-alone method for OSR = 64, and 3–18 dB for OSR = 128. Not shown are simulation results for OSR = 32 where we find a peak SNR for the DFT-alone method of 49 dB. This result is comparable to a result in [20] for an OSR of approximately 36 where a peak SNR of about 50 dB is shown. Comparing Figs. 8 and 18, we find that the gap between the DFT-alone and DFT-POCS curves decreases when the SNR definition is modified to exclude signal frequency errors. This indicates that the DFT-POCS method better suppresses these errors.

Fig. 19 shows SNR results for the fourth-order encoder with a DFT-based band limitation. The results are obtained exactly in the same way as those in Fig. 17, except that the SNR excludes linear errors, and windowing is performed as described in Section IV-E-1. This enables comparison with [8] in which a peak SNR of about 95 dB is reported. In Fig. 19 the peak SNR's are 96 dB for the DFT-alone decoder and 101 dB for the DFT-POCS decoder. The DFT-POCS decoder is 3–10 dB superior to the DFT-alone decoder.

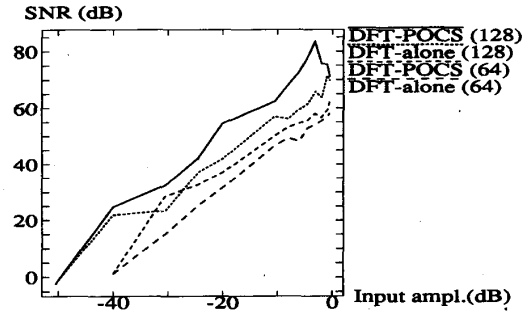

Fig. 18. SNR as a function of input amplitude for the single-loop encoder. The SNR calculation excludes linear errors.

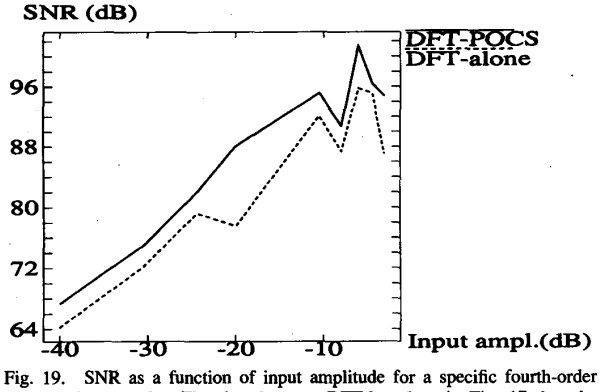

Fig. 19. SNR as a function of input amplitude for a specific fourth-order interpolative encoder. The decoders are DFT-based as in Fig. 17, but the SNR excludes linear errors, and windowing is performed.

Other simulation results show that if we use the SVD-based method for band limitation, and if the SNR is defined to exclude linear errors and to include windowing of the DFT, the peak SNR's are 93 dB for the SVD-alone decoder and 98 dB for the SVD-POCS decoder. Under these conditions, the SVD-POCS decoder is 3–15 dB superior to the SVD-alone decoder.

3) Tone Suppression: In this section we consider the problem of spectral tones that is known to plague single-loop and double-loop encoders [23]–[25]. It has been shown theoretically that in single-loop encoders with constant or sinusoidal inputs, the spectrum of the quantization noise is purely discrete [23], [24]. The part of the noise spectrum that falls in signal baseband gives rise to objectionable tones if a linear low-pass decoder is employed. For the double-loop encoder, a similar behavior has been observed empirically [25]. Spectral tones are known to be absent in multi-stage encoders [26], [27], and it is commonly held that tones are also absent in higher-order interpolative encoders.

As our proposed POCS algorithm is nonlinear, it is conceivable that the algorithm may be able to exploit time-domain information to suppress baseband tones. On the other hand, it is also conceivable that the algorithm may introduce additional nonlinear distortion. In our simulations, we have found neither of these effects to be present. To test the performance of our algorithm, we have carried out simulations for a large number

of constant inputs, as these are known to give rise to tones. We find that the POCS algorithm quickly converges to an input estimate that contains the same tones as the estimate produced by a linear decoder. As the input estimate resulting from the POCS algorithm satisfies both the time-domain and frequency-domain constraints, there is no deterministic way to distinguish it from the actual constant input. We therefore conclude that the presence of tones in single-loop and double-loop modulators is an inherent property of the encoders that cannot be suppressed by decoders.

4) Comparison to Thao and Vetterli's Work: In this section we compare the simulation results of this chapter to those of Thao and Vetterli [11]–[14]. We first discuss the asymptotic performance in terms of SNR versus OSR, and then compare actual SNR levels in dB.

In [12] Thao and Vetterli analytically show the following result: There exists a decoder for an n-stage $\Sigma\Delta$ encoder that statistically achieves an asymptotic SNR versus OSR performance of at least $(2n + 2) \cdot 3$ dB/octave, whereas traditional linear decoding is known to be limited to $(2n+1)\cdot 3$ dB/octave. The result is shown under the assumptions that the quantization noise is white, and that the input signal is chosen at random from the set of band-limited signals. They verify numerically that a POCS algorithm obtains $(2n + 2) \cdot 3$ dB/octave performance.

In Section IV-E we found an SNR versus OSR performance of approximately 12 dB/octave for the single-loop encoder with the DFT-POCS algorithm. Similarly, we found slopes of approximately 18 dB/octave for the double-loop and two-stage encoders with OSR's less than about 80, whereas the slopes appear to slightly decrease for OSR's exceeding 80. The results of 12 dB/octave and 18 dB/octave are in agreement with the results obtained by Thao and Vetterli. However, the decreasing slopes for large OSR's require an explanation.

One possible explanation is that our space of band-limited sinusoidal encoder inputs has dimension N/OSR that is on the order of hundreds. On the other hand, Thao and Vetterli implicitly assume the input frequency to be known, so that the dimension of their input space is 3, namely a dc component and the amplitude and phase of a sinusoidal component. It is possible that knowing the input frequency gives their algorithm an asymptotic advantage. As Thao and Vetterli's theorem is derived under the unrealistic assumption of white quantization noise, the theorem does not show that our algorithm should achieve an asymptotic SNR performance of $(2n + 2) \cdot 3$ dB per octave of oversampling.

Other factors in the simulation set-ups may also play a part in the observed differences. In particular, we consider sinusoidal inputs with a range of amplitudes and no dc component, whereas Thao and Vetterli consider sinusoidal inputs with a fixed amplitude of 0.5, but a random dc component between -0.5 and 0.5. The inclusion of a sizable dc component in the input signal as well as in the SNR calculations may conceivably mask small performance changes in the dynamic behavior. This is because encoders with filter poles at dc achieve relatively better SNR's for dc inputs than for dynamic inputs, as the filter magnifies dc errors with larger gain than it magnifies dynamic errors.

One might speculate that the probable cause of the decreasing slopes for large OSR's is that we keep the sample size N constant at 4096, whereas Thao and Vetterli let N be proportional to the OSR. Clearly, if the sample size is finite, the SNR must also be finite even for infinite OSR, that is, constant signals; it therefore appears that we might be beginning to see the effect of finite N at SNR's of about 85–90 dB. However, we find by simulation that no gain in SNR is obtained by doubling the sample size.

We now discuss the absolute performance of the POCS algorithms, and their performance compared to linear filtering. In our results, we found improvements in SNR over linear filtering for all practical input amplitudes and OSR's. Thao and Vetterli do not report simulation results for linear filtering, but rather use a standard linearized formula to approximate its performance. In [13] they report that for the single-loop encoder, the SNR gain over linear filtering ranges from 0 dB at OSR = 20 to about 6 dB at OSR = 128. For the two-stage encoder, the SNR gains are -3 dB at OSR = 20 and $+4$ dB at OSR = 128, and for the double-loop encoder, the gain ranges from -8 dB at OSR = 20 to -2 dB at OSR = 128. The POCS algorithm is reported to be inferior to linear filtering for OSR's less than about 50 for the two-stage encoder, and for OSR's less than about 256 for the double-loop encoder. The authors attribute the negative SNR gains to the fact that they use the zero signal rather than the quantizer output as their initial input estimate; they justify this choice by stating that it avoids artifacts in the evaluation of the alternating projection algorithm. We have not observed such artifacts in our simulations. Another possible explanation for the reported SNR loss may be that the authors do not simulate the performance of linear filtering, and the theoretical expression for the performance of linear filtering that they use appears to exclude linear errors.

Finally, we compare the actual SNR numbers obtained with our POCS algorithm to those obtained by Thao and Vetterli for the same encoders and input amplitudes. This comparison cannot be made fair, as Thao and Vetterli include a random dc component in their input signal; they also fix the input frequency at the baseband edge, which results in a lower SNR than for input frequencies well within baseband. With these reservations, we find for the single-loop encoder that our SNR is 54 dB, whereas Thao and Vetterli report approximately 46–47 dB. For the two-stage encoder, we find and SNR of 79 dB, whereas Thao and Vetterli report approximately 73 dB. No absolute numbers are reported for the double-loop encoder. We find these numbers to be in acceptable agreement. Without detailed knowledge of both simulations, it does not appear possible to say whether some SNR loss can be attributed to Thao and Vetterli's approximate, but faster time-domain projection.

V. Decoding in the Presence of Non-Idealities

In this section we present simulation results that describe the effects of a number of nonidealities on the performance of the POCS decoding algorithm. The numerical values of the nonidealities are assumed unknown, but can easily be

TABLE II
SENSITIVITY OF THE POCS-DFT DECODER TOWARDS VARIOUS NONIDEALITIES THAT ARE UNKNOWN TO THE DECODER. THE SHOWN NONIDEALITIES RESULT IN A PEAK SNR LOSS OF APPROXIMATELY 2 dB. THE QUANTITIES GIVEN IN PERCENT ARE MEASURED RELATIVE TO THE QUANTIZER STEP SIZE. THE QUANTITIES GIVEN FOR THE LEAK ARE THE MINIMUM ACCEPTABLE OPERATIONAL AMPLIFIER (OP-AMP) GAINS

Encoder	OSR	Nom. SNR	Gain	Leak	State	Quant.	Noise
Single-loop	64	61 dB	N/A	1500	2.5%	2.5%	0.75%
	128	84 dB	N/A	7500	2.0%	0.5%	0.15%
Double-loop	64	79 dB	1%	5000	≥ 15%	10%	0.2%
	128	92 dB	0.1%	50000	≥ 15%	5%	0.01%
Two-stage	64	86 dB	1%	10000	≥ 15%	≥ 15%	0.1%
	128	101 dB	0.005%	150000	5%	2.5%	0.0075%
Fourth-order	48	93 dB	≥ 5%	750	3%	2%	0.01%

compensated for in the algorithm if known. The nonidealities we consider are integrator gain and leak, initial state offset, quantizer offset and input noise, all of which are described in [3]. The encoders we consider are the single-loop, double-loop, two-stage and fourth-order interpolative encoders. The results are summarized in Table II. We first describe the assumptions underlying the results, then present detailed results for the fourth-order encoder, and finally comment on the results summarized in Table II.

The simulation results in Table II are generated using the DFT as the band limitation and considering a single 4 K block of samples. The parameters of the simulations are the same as in Sections IV-A–IV-D. The SNR is calculated in the same way as in those sections, that is, by including linear errors and not performing windowing. The peak SNR is used as a concise statistic to describe the performance of a modulator. For encoders with more than one integrator or quantizer, the nonidealities are assumed for simplicity to be the same for all elements that are nominally identical. The levels of nonidealities reported in Table II result in a peak SNR loss of approximately 2 dB.

We now present detailed results for the fourth-order encoder showing the sensitivity of the POCS decoder as well as the DFT-based linear decoder towards nonidealities. As the encoder contains nine scaling factors, we have not exhaustively graphed the sensitivities towards perturbations in each gain. However, Table III shows the changes in peak SNR resulting from perturbations of ±5% in each gain from its nominal value. We see that in no case does the perturbation result in more than a 2 dB degradation in peak SNR, and in fact the peak SNR of the POCS decoder often increases as a result of a perturbation. Fig. 20 shows that the DFT-POCS decoder requires an op-amp gain of approximately 750, whereas the DFT-alone decoder requires a gain of 300. Fig. 21 shows the sensitivity of the decoders towards initial integrator states. For simplicity, we assume that all state offsets are identical. We see that the DFT-POCS decoder can tolerate an initial state offset of about 3%, whereas the linear decoder is insensitive to initial integrator states. Not shown are figures demonstrating that the DFT-POCS and DFT-alone decoders require a quantizer offset no larger than 2% and 5%, respectively, and that the noise levels should be no larger than 0.01% and 0.03%, respectively.

In general, the results in Table III indicate that the sensitivity increases as the OSR and hence the SNR increases. We also see that for comparable SNR performance, the different encoders mostly have comparable sensitivities. The exception to this rule is the fourth-order interpolative encoder, which exhibits quite low sensitivity towards nonidealities, taking into account its peak SNR of 93 dB. The exact reason for the decreased sensitivity is unclear, but it appears reasonable to conjecture that it is related to the pole locations of the fourth-order encoder. Specifically, the single-loop, double-loop and two-stage encoders all have their poles at dc, whereas the fourth-order encoder has its poles at non-dc frequencies within signal baseband.

TABLE III
PEAK SNR CHANGES IN dB FOR THE FOURTH-ORDER INTERPOLATIVE ENCODER AS A FUNCTION OF PERTURBATIONS IN OPEN-LOOP FILTER GAINS

	POCS-DFT		DFT-alone	
Coeff.	+5%	−5%	+5%	−5%
A_0	+0.39	+1.43	−0.69	−0.11
A_1	−0.06	+3.59	−0.12	−0.27
A_2	−1.06	+3.43	0.00	+0.07
A_3	+1.60	+3.54	+0.21	−0.30
A_4	+1.46	+1.65	+0.26	−0.44
B_1	+1.99	+3.76	+0.20	0.00
B_2	−1.86	+0.11	+0.58	−0.73
B_3	+2.64	+3.28	−0.17	+0.06
B_4	+1.83	+3.06	−1.15	+0.84

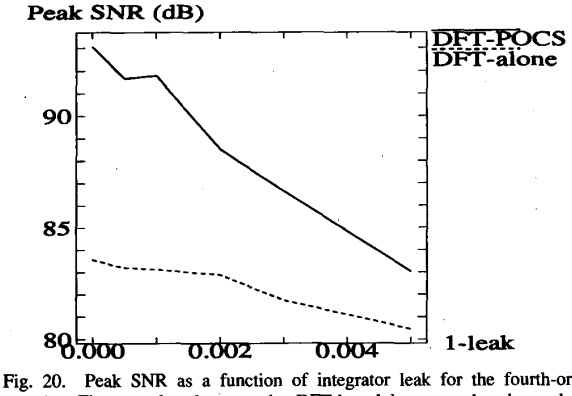

Fig. 20. Peak SNR as a function of integrator leak for the fourth-order encoder. The two decoders are the DFT-based low-pass decoder and the DFT-based POCS decoder.

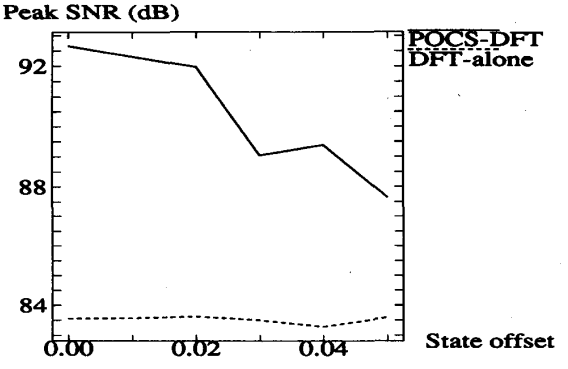

Fig. 21. Peak SNR as a function of integrator state offsets for the fourth-order encoder. The two decoders are the DFT-based low-pass decoder and the DFT-based POCS decoder.

VI. Conclusion

We have shown by simulations that for four representative $\Sigma\Delta$ modulator architectures, significant gains in SNR can be achieved using sophisticated decoding techniques rather than linear decoding. The results are valid both for block decoding and, more realistically, for sliding-block decoding.

The presented decoding method is computationally complex. However, our results can serve as upper bounds on the achievable SNR, and would be useful for future research in simplified, nonlinear decoding algorithms. Such research would also address the sensitivity of the algorithm towards circuit imperfections.

Appendix
Quadratic Programming Approximation

We first rewrite (2) by partitioning all vectors and matrices into blocks of L samples, where L divides N. Specifically, we set $P = N/L$ and

$$\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_P\}$$
$$\mathbf{y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_P\}$$
$$\mathbf{z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_P\} \triangleq \mathbf{Zs}$$
$$\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_1, \cdots, \mathbf{Q}_P\}$$
$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{H}_2 & \mathbf{H}_1 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{H}_{P-1} & \mathbf{H}_{P-2} & \cdots & \mathbf{H}_1 & \mathbf{0} \\ \mathbf{H}_P & \mathbf{H}_{P-1} & \cdots & \mathbf{H}_2 & \mathbf{H}_1 \end{bmatrix}.$$

Then (2) can be written

$$\mathbf{Q}_n \sum_{i=1}^{n} \mathbf{H}_{n-i+1}\mathbf{x}_i \le \mathbf{Q}_n \left[\sum_{i=1}^{n} \mathbf{H}_{n-i+1}\mathbf{y}_i - \mathbf{z}_n \right], \quad n = 1, 2, \cdots, P \quad (5)$$

and the objective is to find the sequence $\hat{\mathbf{x}}$ that satisfies (5) and minimizes

$$d^2 = \sum_{n=1}^{P} \| \hat{\mathbf{x}}_n - \mathbf{x}_n \|_2^2.$$

We can solve this N-dimensional QP problem approximately by solving the following N/L L-dimensional QP problems in sequence: Minimize $\| \hat{\mathbf{x}}_n - \mathbf{x}_n \|_2^2$ subject to

$$\mathbf{Q}_n\mathbf{H}_1\mathbf{x}_n \le \mathbf{Q}_n \left[\sum_{i=1}^{n-1} \mathbf{H}_{n-i+1}(\mathbf{y}_i - \hat{\mathbf{x}}_i) + \mathbf{H}_1\mathbf{y}_n - \mathbf{z}_n \right], \quad n = 1, 2, \cdots, P.$$

To avoid large errors at the beginnings of L-dimensional QP subblocks, we can overlap the QP subblocks so that each subblock continues into the immediately following subblock. Specifically, we denote the length of the nth QP subblock by $\lambda(n)$, where

$$\lambda(n) \triangleq \begin{cases} L & \text{for } n \ne N/L \\ \ell & \text{for } n = N/L \end{cases}$$

and we define the index set over which the nth QP problem is solved by

$$\Lambda(n) \triangleq \{(n-1)\ell, \cdots, (n-1)\ell + \lambda(n) - 1\}.$$

We require that the length ℓ divides N, and we solve N/ℓ QP problems of which all but one have dimension L, and the last one has dimension ℓ. The last one is shorter than the others because we only have N and not $N + L - \ell$ samples available. We use the index set $\Lambda(n)$ as a subscript of sequences to denote the index limitation to $\Lambda(n)$; for instance, $\mathbf{x}_{\Lambda(n)} = \{x_{(n-1)\ell}, \cdots, x_{(n-1)\ell+\lambda(n)-1}\}$. We also use $\Lambda(n)$ as a subscript for matrices to denote the index limitation in both dimensions to $\Lambda(n)$; for instance, $\mathbf{Q}_{\Lambda(n)} = \operatorname{diag}(y_{\Lambda(n)}) = \{y_{(n-1)\ell}, \cdots, y_{(n-1)\ell+\lambda(n)-1}\}$. From the solution of each QP subproblem, we only keep the ℓ first samples, that is, we set $\hat{\mathbf{x}}_n$ equal to the first ℓ samples of $\mathbf{x}_{\Lambda(n)}$. We thus solve the following N/ℓ QP problems in sequence: Minimize $\| \hat{\mathbf{x}}_{\Lambda(n)} - \mathbf{x}_{\Lambda(n)} \|_2^2$ subject to

$$\mathbf{Q}_{\Lambda(n)}\mathbf{H}_{\Lambda(n)}\mathbf{x}_{\Lambda(n)} \le \mathbf{Q}_{\Lambda(n)} \left[\sum_{i=1}^{n-1} \mathbf{H}_{n-i+1}(\mathbf{y}_i - \hat{\mathbf{x}}_i) + \mathbf{H}_{\Lambda(n)}\mathbf{y}_{\Lambda(n)} - \mathbf{z}_{\Lambda(n)} \right], \quad n = 1, 2, \cdots, N/\ell$$

Acknowledgment

The authors thank S. Nadeem and Prof. B. N. Parlett for useful discussions.

References

[1] R. M. Gray, "Oversampled Sigma-Delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481–488, May 1987.
[2] P. R. Gray, D. Pederson, and A. Zakhor, "National Science Foundation proposal," Univ. California at Berkeley, Oct. 1988.
[3] S. Hein and A. Zakhor, "Optimal decoding for data acquisition applications of Sigma Delta modulators," *IEEE Trans. Signal Processing*, vol. 41. pp. 602–616, Feb. 1993.
[4] S. Hein and A. Zakhor, "Optimal decoding for data acquisition applications of Sigma Delta modulators," U.S. Patent No. 5164727, Nov. 17, 1992.
[5] S. Hein and A. Zakhor, "Reconstruction of oversampled band-limited signals from Sigma Delta encoded binary sequences," in *Proc. Int. Conf. Acoust, Speech, Signal Processing* (San Francisco, CA), Mar. 1992.

[6] S. Hein and A. Zakhor, "Iterative reconstruction of Sigma Delta encoded signals," in *Proc. Int. Conf. Indust., Appl. Math.* (Washingtion, DC), July 1991, p. 88.
[7] J. C. Candy, "A use of double integration in Sigma-Delta modulation," *IEEE Trans. Commun.*, vol. COM-33, pp. 249–258, Mar. 1985.
[8] K. C. 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.
[9] D. C. Youla and H. Webb, "Image restoration by the method of convex projections: Part I—theory," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 81–94, Oct. 1982.
[10] M. I. Sezan and H. Stark, "Image restoration by the method of convex projections: Part II—applications and numerical results," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 95–101, Oct. 1982.
[11] N. T. Thao and M. Vetterli, "Oversampled A/D conversion using alternate projections," in *Proc. Twenty-Fifth Annu. Conf. Inform. Sci., Syst.*, 1991, pp. 241–248.
[12] N. T. Thao and M. Vetterli, "Convex coders and oversampled A/D conversion: Theory and algorithms," CU/CTR/TR 289-91-81, Columbia Univ., Dec. 16, 1991.
[13] N. T. Thao and M. Vetterli, "Optimal MSE signal reconstruction in oversampled A/D conversion, part I: Deterministic analysis of A/D conversion and convexity, part II: Algorithms and numerical experiments," submitted to *IEEE Trans. Signal Processing*, Mar. 13, 1992.
[14] N. T. Thao and M. Vetterli, "Optimal MSE signal reconstruction in oversampled A/D conversion using convexity," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992.
[15] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.
[16] S. Hein and A. Zakhor, "Theoretical and numerical aspects of an SVD-based method for band-limiting finite extent sequences," *IEEE Trans. Signal Processing*, vol. 42, no. 5, May 1994.
[17] E. J. Diethorn and D. C. Munson, Jr., "A linear time-varying system framework for noniterative discrete-time band-limited signal extrapolation," *IEEE Trans. Signal Processing*, vol. 39, pp. 55–68, Jan. 1991.
[18] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis, and uncertainly—v: The discrete case," *Bell Syst. Tech. J.*, vol. 57, pp. 1371–1430, May/June 1978.
[19] B. N. Parlett, *The symmetric eigenvalue problem.* Englewood Cliffs, NJ: Prentice-Hall, 1980.
[20] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from Sigma-Delta modulation," *IEEE Trans. Commun.*, vol. COM-29, pp. 1316–1323, Sept. 1981.
[21] B. E. Boser, "Design and implementation of oversampled analog-to-digital converters," Ph.D. dissertation, Stanford Univ., Stanford, CA, Oct. 1988.
[22] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1989.
[23] R. M. Gray, "Spectral analysis of quantization noise in a single-loop Sigma-Delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588–599, June 1989.
[24] R. M. Gray, W. Chou, and P. W. Wong, "Quantization noise in single-loop Sigma-Delta modulation with sinusoidal inputs," *IEEE Trans. Commun.*, vol. 37, pp. 956–968, Sept. 1989.
[25] B. E. Boser and B. A. Wooley, "Quantization error spectrum of $\Sigma\Delta$ modulators," in *Proc. Int. Symp. Circuits, Syst.*, 1988.
[26] W. Chou, P. W. Wong, and R. M. Gray, "Multistage Sigma-Delta modulation," *IEEE Trans. Inform. Theory*, vol. 35, pp. 784–796, July 1989.
[27] P. W. Wong and R. M. Gray, "Two-stage Sigma-Delta modulation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1937–1952, Nov. 1990.

Søren Hein (S'88–M'93) was born in May 1968 in Copenhagen, Denmark. He received the M.Sc. degree in electrical engineering from the Technical University of Denmark in January 1989 and the Ph.D. degree in electrical engineering from the University of California at Berkeley in May 1992.

His research interests include algorithmic aspects of oversampled A/D conversion, signal reconstruction, and signal and image processing for medical and other applications. He has also worked on error-correction coding for satellite communications.

Avideh Zakhor received the B.S. degree from the California Institute of Technology, Pasadena, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering, in 1983, 1985, and 1987, respectively.

In 1988, she joined the faculty at the University of California at Berkeley, where she is currently Assistant Professor in the Department of Electrical Engineering and Computer Sciences. Her research interests are in biomedical data and in the general area of signal processing and its applications to images and video. She has been a consultant to a number of industrial organizations and holds four U.S. patents.

Ms. Zakhor was a General Motors scholar from 1982 to 1983, received the Henry Ford Engineering Award and Caltech Prize in 1983, was a Hertz Fellow from 1984 to 1988, received the Presidential Young Investigators (PYI) award, IBM junior faculty development award, and Analog Devices junior faculty development award in 1990, and Office of Naval Research (ONR) young investigator award in 1992. She is currently Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.