Variable-Structure Compensation of Delta–Sigma Modulators: Stability and Performance

Takis Zourntos, *Member, IEEE,* and David A. Johns, *Fellow, IEEE*

***Abstract*—We develop a compensation method for continuous-time delta–sigma modulators valid for loop filters of arbitrary order. Our approach, based on variable-structure theory, accommodates multilevel quantization and dithering. Stability is rigorously proved under the assumption of infinite sampling rate and is accompanied by an analytic characterization of performance. A slight modification of the basic compensator provides a defence against parametric uncertainty through the use of variable-integrator damping.**

***Index Terms*—Analog, continuous-time filtering, delta–sigma modulation, electronics, integrated circuits, sliding-mode, stability, variable-structure control.**

I. Introduction

THE BANDWIDTH requirements of emerging communication standards have prompted interest in the development of data converter technologies. Although traditionally confined to low-speed applications such as audio-range signal processing and narrow-band communications, recent efforts have demonstrated the feasibility of delta–sigma modulation techniques for wide-band data conversion [1], [2]. A current trend in area-efficient single-loop modulator integrated circuits operating at high sampling rates is the use of continuous-time loop filters in combination with relatively coarse (often single-bit) quantizers [3], [4].

In this paper, we develop compensation strategies suitable for continuous-time modulators employing loop filters of arbitrary order. This work is of both practical and academic interest. Our approach is based on the use of variable-structure techniques (for an introduction see [5] or [6]) which have received some, though not extensive, attention in the design of analog electronic systems [7], [8]. Our methods provide two main benefits. First, a "soft-reset" effect, i.e., stabilization with potentially less degradation in signal-to-noise-ratio (SNR) than that of conventional reset-compensation. Second, variable-integrator damping to yield robust performance in the face of uncertainties in filter components.

A. Background

The basic delta–sigma modulator architecture (Fig. 1) is *stable* if the states of the loop filter are bounded and the input to the quantizer is within specified limits, given any initial condition within a subset of state space and any input signal within certain bounds. Definition 1 (Section III) provides a formal statement of modulator stability.

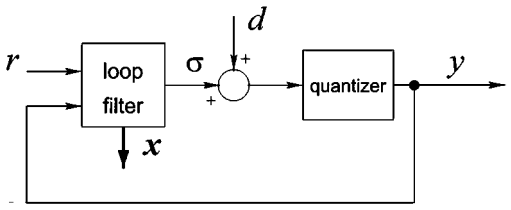

Fig. 1. A basic delta–sigma modulator with dither input, d. The *quantizer* element is a clocked device, and the *loop filter*, with state vector $\boldsymbol{x}$ (vector signals are denoted by bold lines in all figures), is a linear time-invariant (LTI) system.

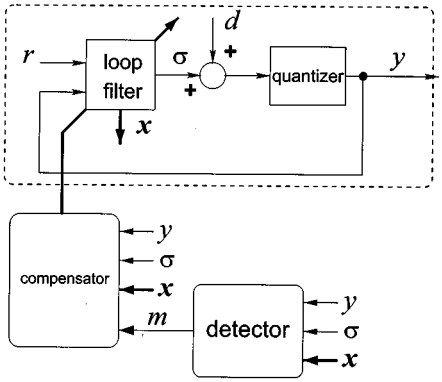

Fig. 2. The compensated delta–sigma modulator contains the basic modulator (within the dashed boundary) in addition to a set of stabilizing elements. The *detector* block indicates the onset of large states by employing a user-defined metric. The signal m prompts the *compensator* to return the system to desirable operating conditions with a specific type of corrective action. The loop filter is not LTI if its parameters or state are alterable through the action of the compensator.

With the "accumulation error condition" described in [9], or the invariant-set results of Schreier *et al.* [10], it is possible to construct basic delta–sigma modulators which are stable. However, the accumulation error condition relies on the use of multilevel quantization and may impose significant costs in terms of area or power. For arbitrary inputs, the work of Schreier *et al.* is only valid for modulators up to second order and therefore may have limited utility.

A rule-of-thumb approach is often used to design modulators for which the symptoms of instability arise less frequently [11]. Nonetheless, the thrust to improve performance has made instability in the basic delta–sigma modulator practically unavoidable. The onset of large states is typically handled on a contingency basis as indicated in Fig. 2. In the case of a resetting single-bit modulator, the loop filter is equipped with reset switches on each integrator, the *detector* may simply count the

Manuscript received August 21, 2000; revised June 6, 2001 and July 16, 2001. This paper was recommended by Associate Editor M. J. Ogorzalek.

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada.

Publisher Item Identifier S 1057-7122(02)00279-9.

1057–7122/02$17.00 © 2002 IEEE

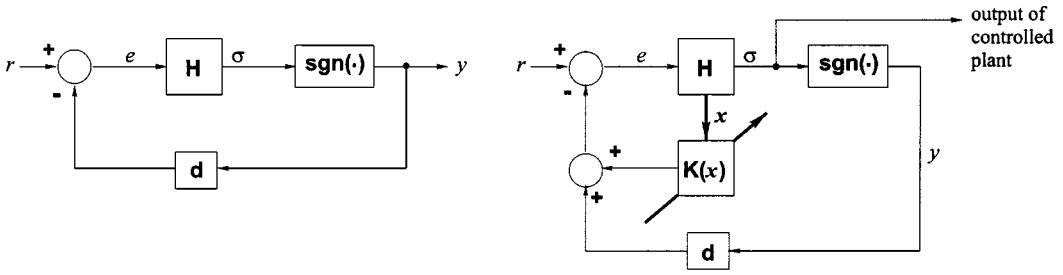

Fig. 3. Architectural comparison of single-bit interpolative modulator (left) and variable-structure control system (right); the corresponding signals have identical labels. The modulator only lacks the stabilizing switching feedback, $\mathbf{K}(\boldsymbol{x})$. The block H denotes a linear time-invariant system and δ represents a constant gain.

number of consecutive "1" s or "0" s (the so-called *run-length*) in y to estimate the degree of quantizer overload, and the *compensator* triggers the integrator reset switches to close or open via the signal ξ. Other stabilization strategies are also amenable to this general form. To avoid confusion, we refer to the system shown in the figure, including the basic delta–sigma modulator, as a *compensated modulator*. Therefore, the complete architecture may be stable, in spite of the fact that the basic modulator, taken alone, is not.

B. Compensation Strategies

The simplest means of ensuring that the states of the compensated modulator are within desired ranges is to reset all integrators if a threshold is exceeded. This guarantees stability but can adversely impact performance. In the limit, as the frequency of resets tends to infinity, we observe a noise transfer function of unity, and a signal transfer function of zero. Thus, while resetting provides stability, it can, in principle, also yield the worst possible resolution.

Various compensation strategies have been proposed to improve robustness without significantly degrading performance. State-limiting strategies which attempt to confine integrator outputs to "stable regions" of state-space as suggested in ([11], Section 4.6) may be difficult to implement since such regions are not known for arbitrary inputs. The approach of Moussavi and Leung guarantees stability for discrete-time modulators using a local-feedback strategy and digital partial-cancellation of stabilizing signals [12]. However, the validity of the technique in the case of continuous-time loop filters is not proved.

In this paper, we apply *variable-structure* methods in the development of stable compensated modulators. Variable-structure theory is a branch of systems control in which switching elements are used for stabilization and tracking. Our approach may be viewed as a "natural choice" for modulator stabilization for the following reasons. First, as shown in Fig. 3, the interpolative delta–sigma modulator and the variable-structure control system have similar form. Second, variable-structure control algorithms only require switching elements and fixed-gain amplifiers and can therefore be implemented inexpensively. Third, variable-structure methods are suitable for systems with discontinuous dynamics so that analytically proving stability for arbitrary-order loop filters is possible.

II. Overview of Proposed Compensators

Each of the architectures presented in this section is based on the compensated modulator form of Fig. 4. Throughout the paper we assume that the *nominal* loop filter, denoted by H_0, has the controllable-canonical form realization shown in Fig. 5 (the state-model for H_0 is assumed to be given by $(\mathbf{A}_0, \boldsymbol{b}_0, \boldsymbol{c}_0, 0)$, described in detail in Section 3). This "direct" form is often used in continuous-time implementations ([13], [14]). Although other realizations are amenable to the formal methods applied in this work, they are not considered to simplify our presentation. We augment H_0 with switching feedback elements k_i to obtain the effective loop filter H shown in Fig. 6. These gains are distributed across the filter H to help minimize the effects of nonidealities on modulator resolution.

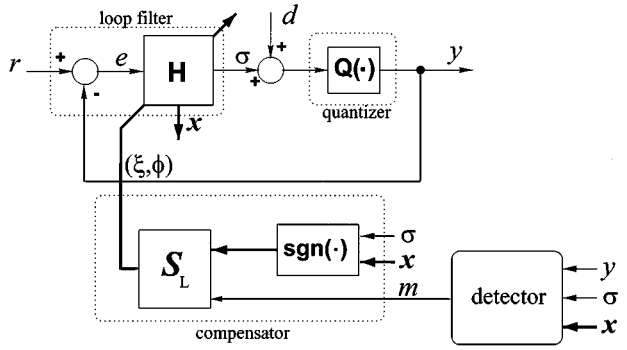

Fig. 4. The general form of our proposed modulator compensation architectures. The quantizer $Q(\cdot)$ may be multi-level. The block S_L determines appropriate settings for the k_i gains (Figs. 7 and 12) based on the sign of σ and the signs of loop filter states. An estimate of the magnitude of the state vector $\boldsymbol{x}$ enables the use of operating *modes* described in Section II-C.

A. Soft-Resetting (SR) Compensator

The first contribution of this paper is a stabilization method best described as a mild form of resetting. Once activated, our "soft-reset" can ensure that from any initial condition within a specified set: 1) quantizer overload is avoided; 2) states are bounded; and 3) states enter a neighborhood of the origin and remain until compensation is deactivated. As long as the input signal is within its predetermined bounds, it does not affect this process. In addition, as with conventional reset-compensation, soft-resetting can guarantee that loop filter states substantially *shrink*, providing a defence against oscillatory instabilities which may not be as effectively countered by other stabilization techniques. Unlike conventional resetting, soft-resetting yields appreciable modulator performance under permanent activation (this property is demonstrated in Section II-B).

The *detector* (shown in Fig. 4) outputs $m = 0$ under normal operating conditions and sets $m = 1$ if corrective measures are required. The implementation details of this block are left to

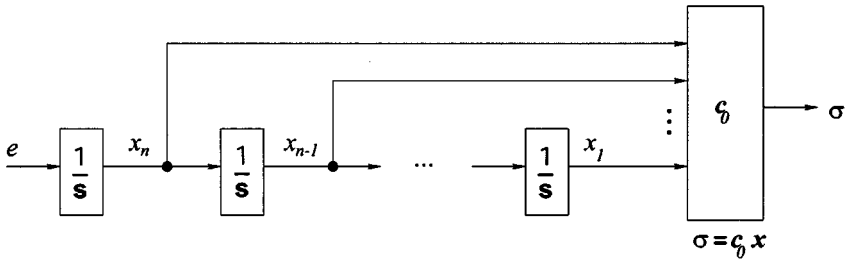

Fig. 5. Nominal loop filter structure, based on controllable-canonical form.

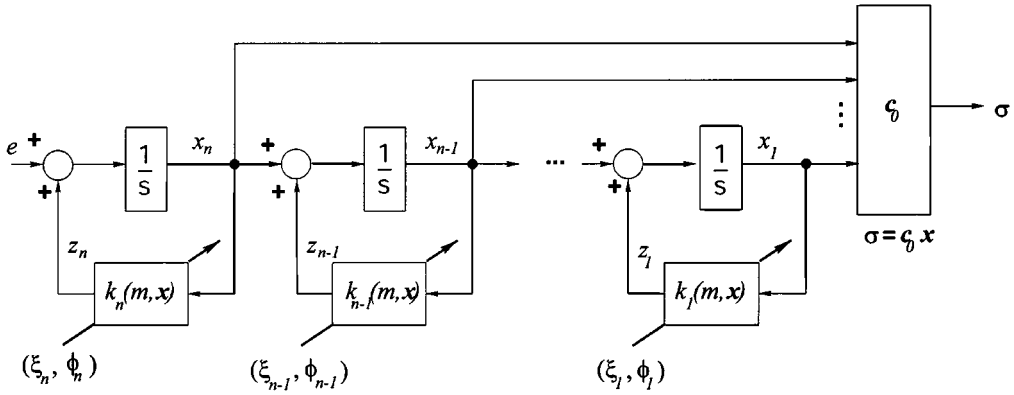

Fig. 6. Effective loop filter structure for use with compensation strategies proposed in this paper. Switching feedback elements k_i are shown in Figs. 7 and 12.

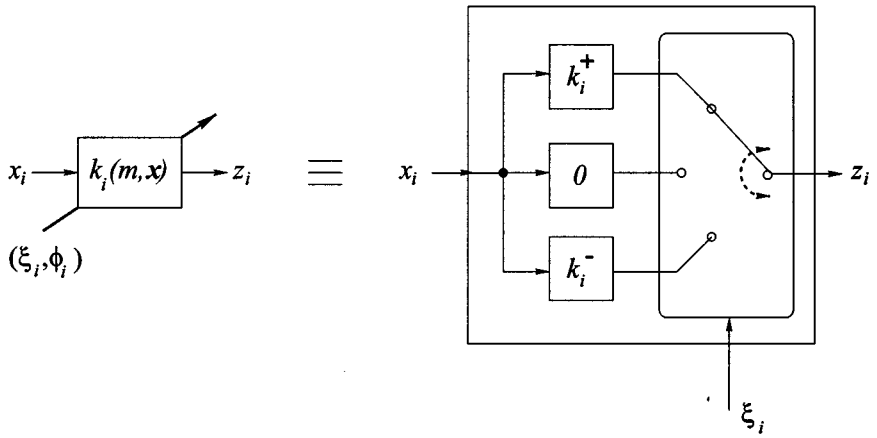

Fig. 7. The switchable feedback gain element for the soft-resetting compensator. It is defined as $k_i(m, \boldsymbol{x}) = \bar{k}_i + \varepsilon_i \operatorname{sgn}(\sigma) \operatorname{sgn}(x_i)$, where $\bar{k}_i$ is a real constant and $\varepsilon_i > 0$. Thus $k_i^+ = \bar{k}_i + \varepsilon_i$ and $k_i^- = \bar{k}_i - \varepsilon_i$. The zero-gain switch position corresponds to no compensation. The parameter ε_i is needed to counter parametric uncertainties and to speed convergence of $\boldsymbol{x}(t)$.

the designer. Switching is based on specific rules expressed in the switching logic block, $\mathbf{S}_L$, which only requires the signs of the states x_i and the sign of σ. The damping element k_i $(i = 1, \ldots, n)$ is illustrated in Fig. 7.

The functioning of the SR-compensated modulator can be explained with the aid of Fig. 8. Once the SR compensator is triggered, the loop filter state vector moves such that the magnitude of the loop filter output decreases monotonically. The states eventually enter the set $\boldsymbol{\Omega}_0$ (a closed neighborhood of the origin) and remain as long as $m = 1$. Throughout this process, all loop filter states are bounded. Operation as a basic modulator resumes when m is cleared to 0. Note that the compensator may be triggered from any point (initial condition) in the state space since the soft-reset is globally stabilizing.

B. Performance of SR-Compensated Modulator

We now develop analytic SNR formulas to estimate the performance of the SR-compensated modulator. In the following discussion, we assume that soft-resetting is permanently on. Our results therefore cannot determine the SNR loss from a transient application of soft-resetting, but are relevant to long-term activation of the SR compensator. The analysis provides us with insight into the SR compensator from a signal-processing perspective. Simulations support our theoretical findings. Model parameters, including noise transfer function (NTF) out-of-band-gain (OOBG) and oversampling ratio (OSR) settings, are provided for reference in Table I and the performance of compensation methods in Table 2.

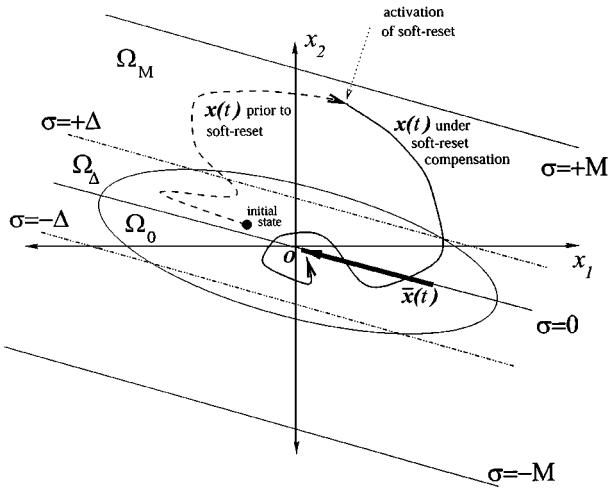

Fig. 8. State-space depiction of soft-resetting action, second-order case shown. The hyper-plane $\sigma = 0$ corresponds to a loop filter output of 0. This set is contained within a region $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ whose width, $\boldsymbol{\Delta}$, is described in the Formal Development Section. The path of the system state, $\boldsymbol{x}(t)$, is illustrated both before and after the activation of the soft-reset. The planes $\sigma = \pm M$ correspond to quantizer overload limits. The bold trajectory, $\bar{\boldsymbol{x}}(t)$, represents the solution to (49), known as the "sliding-mode" solution, initialized at the point on $\sigma = 0$ nearest to $\boldsymbol{x}(t)$ at the time $\boldsymbol{x}(\cdot)$ intersects $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. As shown in Section III, $\boldsymbol{x}(t)$ must enter and remain within a neighborhood of the origin (contained within $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$), denoted here by $\boldsymbol{\Omega}_0$.

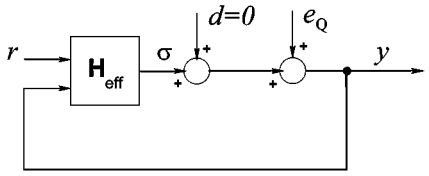

Fig. 9. Linear approximation to SR-compensated delta–sigma modulator intended for performance analysis. It is assumed that m is permanently held at 1 and that the effect of dither can be ignored.

We proceed with the linear approximation to the SR-compensated modulator of Fig. 9. The loop filter H_{eff} is obtained by fixing the *detector* output of the SR-compensator m to 1, and by setting the ε_i (discussed in the caption of Fig. 7) all equal to 0. We assume that quantization noise is described by a uniformly distributed zero-mean stochastic process with variance $\sigma^2_{e_Q}$ which can be shown to equal $(\Delta_Q^2)/12$ where Δ_Q denotes the spacing between adjacent quantization levels. As shown in Appendix A, the effective noise-transfer function of the SR-compensated delta–sigma modulator is given by

$$N_G(s) = \frac{s}{s + c_{0n}} \tag{1}$$

regardless of the number of integrators. Thus, unlike conventional resetting, soft-resetting provides spectral noise-shaping (albeit first order!). The c_{0n} variable denotes the nth entry of the $\boldsymbol{c}_0$ matrix. We obtain the quantization noise power spectral density (PSD)

$$S_{e_Q}(j\Omega) = \int_{-\infty}^{\infty} \sigma^2_{e_Q} \delta_I(t) e^{-j\Omega t} dt = \sigma^2_{e_Q} \tag{2}$$

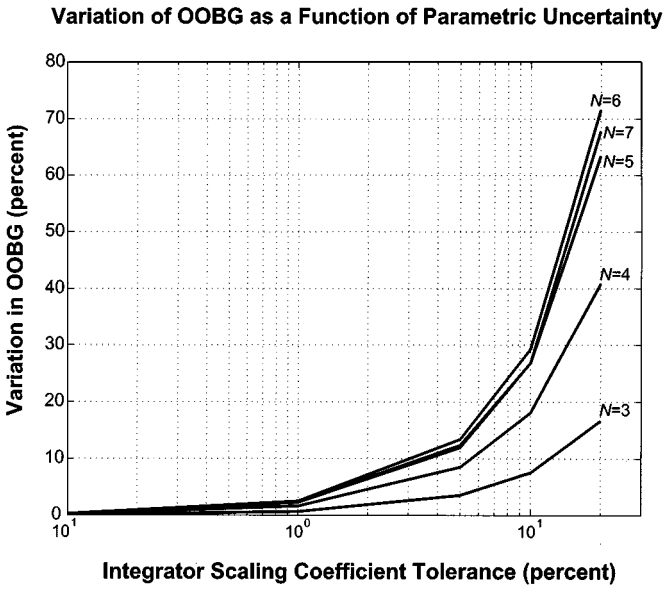

Fig. 10. Variations in NTF OOBG of delta–sigma modulators for lowpass continuous-time loop filters of order N based on integrator scaling coefficient uncertainty. The loop filter is realized in controllable-canonical-form. Device matching is assumed and deviations of up to 20% are applied. The entries of the $\boldsymbol{c}_0$-matrix are free of error.

where $\delta_I(\cdot)$ represents the Dirac delta function. Hence the shaped quantization-noise PSD is given by

$$\begin{aligned} S_N(j\Omega) &= |N_G(j\Omega)|^2 S_{e_Q}(j\Omega) \\ &= \frac{\left(1 + c_{0n}^2\right) \Delta_Q^2}{12 c_{0n}^4} \Omega^2. \end{aligned} \tag{3}$$

Integrating, we estimate the shaped quantization noise power as

$$P_N = \int_{-\Omega_B}^{\Omega_B} S_N(j\Omega) d\Omega = \frac{\left(1 + c_{0n}^2\right) \Delta_Q^2}{18 c_{0n}^4} \Omega_B^3. \tag{4}$$

Assuming a sinusoidal input signal with amplitude α_1 and frequency $\Omega_1 \in (-\Omega_B, +\Omega_B)$ (and so an in-band signal power of $P_S = (\alpha_1^2/2)$), we may express the SNR of the SR-compensated modulator as

$$\begin{aligned} \Gamma_{\text{SR}} &= 10 \log_{10} \left(\frac{P_S}{P_N} \right) \\ &= 10 \log_{10} \left(\frac{9 \alpha_1^2 c_{0n}^4}{\left(1 + c_{0n}^2\right) \Delta_Q^2 \Omega_B^3} \right). \end{aligned} \tag{5}$$

From (5), it is evident that more aggressive nominal NTFs should yield higher SNRs under soft-resetting because c_{0n} tends to increase with OOBG. Since $\Omega_B = (\Omega_S / 2\,\text{OSR})$ (in which Ω_S denotes sampling frequency), resolution should also increase with oversampling ratio. In addition, the use of multi-bit quantization, rather than single-bit quantization, should also improve SNR. Typical values for these parameters (assuming single-bit quantization) suggest that Γ_{SR} should be somewhere in the range of 4- to 8-bits.

TABLE I
SYSTEM AND SIMULATION PARAMETER SETTINGS FOR THE FIFTH-ORDER SR-COMPENSATED MODULATORS USED IN Section II-B. ALL NTF OOBGs REFER TO THOSE OF DISCRETE-TIME PROTOTYPE FILTERS. NOTE THAT δ DEFINES THE QUANTIZER OUTPUT LIMITS $\{+\delta, -\delta\}$. A SINGLE-BIT QUANTIZER IS USED IN EACH SIMULATION

Parameters for Simulated SR-Compensated Modulators							
System Parameters				*Simulation Parameters*			
$H_0(s)$	NTF OOBG	δ	$\{\varepsilon_i\}_{i=1}^n$	record length	OSR	sampling interval	ODE solver
$0.01\frac{67s^4+25s^3+6.0s^2+0.89s+0.066}{s^5}$	1.5	1	$\{0,\ldots,0\}$	65536	(64, 512)	1	4-th-order Runge-Kutta (step size: 0.005)
$0.01\frac{102s^4+62s^3+24s^2+6.0s+0.74}{s^5}$	2.0						
$0.01\frac{139s^4+122s^3+71s^2+27s+5.2}{s^5}$	3.0						

TABLE II
PERFORMANCE OF COMPENSATION METHODS

Performance of Compensation Methods				
OSR	NTF OOBG	SNR with Soft-Reset (dB)		SNR with Conventional Reset (dB)
		predicted	simulated (with dithering)	simulated
64	1.5 (c_{0n}=0.67)	32.3	26.9	4.5
	2.0 (c_{0n}=1.0)	37.8	29.2	3.0
	3.0 (c_{0n}=1.4)	42.0	34.4	2.3
128	1.5 (c_{0n}=0.67)	42.0	39.3	3.2
	2.0 (c_{0n}=1.0)	48.0	39.6	2.9
	3.0 (c_{0n}=1.4)	52.0	40.7	2.1
256	1.5 (c_{0n}=0.67)	51.8	42.3	3.6
	2.0 (c_{0n}=1.0)	57.4	47.5	2.8
	3.0 (c_{0n}=1.4)	61.5	48.5	2.8

C. Variable-Integrator Damping and Soft-Resetting (VIDSR) Compensator

Figs. 10 and 11 illustrate that significant variations in the NTF OOBG can result from small deviations in $\boldsymbol{c}_0$-matrix entries or errors in integrator scaling coefficients (in an actual electronic circuit, *scaled* integrators of the form $(1/s\tau)$ are used; for continuous-time filters, the τ parameter is often a function of a transconductance g_m, and a capacitance C). In general, sensitivity increases with loop filter order. For a typical integrated circuit technology, tolerances in g_m/C ratios without tuning can be more than 30%; such parametic uncertainty can have disastrous effects on the stability of practical implementations. For example, for a seventh-order modulator, if the OOBG changes from 1.5 to 1.7 (an increase of about 13%), the stable input range reduces by more than 50% (determined from simulation).

By using modified switching elements as shown in Fig. 12 and adopting a slightly more complex detection scheme, the soft-resetting compensator can be generalized into the VIDSR compensator. That is, in addition to the soft-resetting effect, the VIDSR compensator permits the adjustment of the system NTF; setting each Δ_i to a small positive value perturbs each pole of the loop filter, and hence each zero of the NTF, slightly away from dc (along the real-axis). It is therefore possible to "tune down" the NTF OOBG from its nominal setting to a lower value and thus counter parametric uncertainty without the need to adjust integrator scaling coefficients or $\boldsymbol{c}_0$ parameters. Although variable-integrator damping can be employed independently of soft-reset compensation (indeed, it may be used in conjunction with conventional resetting), our point is that we incur only a small cost in adjusting the SR compensator to accommodate

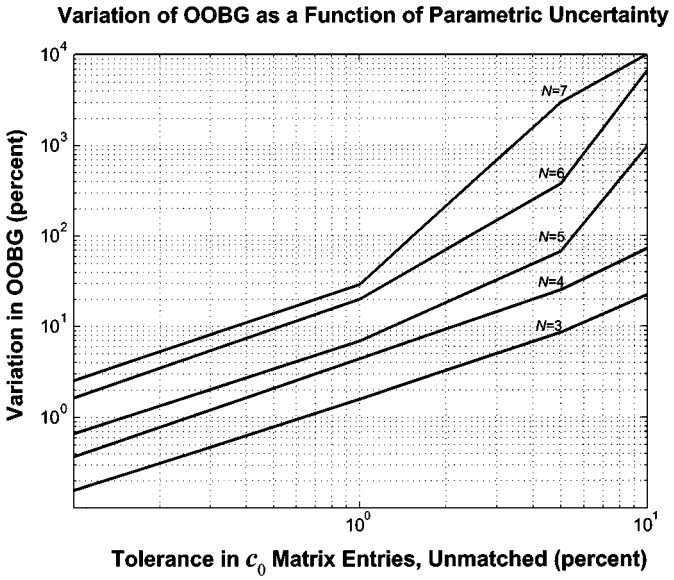

Fig. 11. Variations in NTF OOBG of delta–sigma modulators for lowpass continuous-time loop filters of order N based on mismatching uncertainty in elements of the $\boldsymbol{c}_0$-matrix. Peak mismatch errors of up to 10% are applied. The loop filter is realized in controllable-canonical-form. Integrator scaling coefficients are free of error.

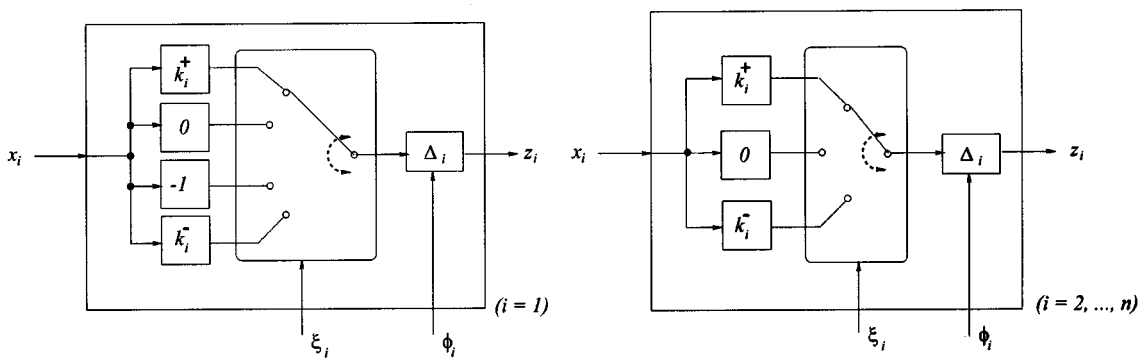

Fig. 12. Modified switchable feedback gain elements to accommodate variable-integrator damping. The Δ_i blocks denote attenuators. The -1 gain (for $i = 1$) enables us to "weight" the control in favor of integrators closer to the quantizer and thus minimize the effects of nonidealities at the modulator output. This additional gain is required because k_1^+ and k_1^- have values which are potentially close to zero [see (15)].

TABLE III
SYSTEM AND SIMULATION PARAMETER SETTINGS FOR THE SEVENTH-ORDER VIDSR-COMPENSATED MODULATOR USED IN Section II. NOTE THAT δ DEFINES THE QUANTIZER OUTPUT LIMITS $\{+\delta, -\delta\}$ (A SINGLE-BIT QUANTIZER IS USED FOR SIMULATION)

Parameters for Simulated VIDSR-Compensated Modulator							
System Parameters				*Simulation Parameters*			
$H_0(s)$ (nominal)	NTF OOBG	δ	$\{\varepsilon_i\}_{i=1}^n$	record length	OSR	sampling interval	ODE solver
$0.01\frac{87s^6+44s^5+15s^4+3.5s^3+0.69s^2+0.064s+0.0034}{s^7}$	1.81	1	$\{0.2, 0.02, 0.02, 0.03, 0.05, 0.07, 0.1\}$	65536	64	1	4-th-order Runge-Kutta (step size = 0.005)

TABLE IV
VARIABLES Δc_{0i} AND Δa_{ij} REPRESENT VARIATIONS IN THE NON-ZERO ENTRIES OF THE $\boldsymbol{c}_0$ AND $\mathbf{A}_0$ MATRICES, RESPECTIVELY, OF THE MODEL OF Table III. THE ACTUAL PARAMETER VALUES ARE OBTAINED THROUGH ITERATIVE RANDOM SEARCHES ASSUMING THAT VARIATIONS ARE UNIFORMLY DISTRIBUTED RANDOM VARIABLES. THE SEARCHES MAXIMIZE THE NTF OOBGs FOR $m = 1$. AS IS CONVENTIONALLY KNOWN, THE NTF OOBG SHOULD BE, IN GENERAL, LESS THAN 1.8 TO OBTAIN A SUFFICIENTLY WIDE "STABLE INPUT RANGE." THUS, AS CAN BE SEEN FROM THE TABLE, THE VARIABLE-INTEGRATOR DAMPING FEATURE OF OUR COMPENSATOR PERMITS TUNING OF THE NTF IN SPITE OF LARGE PARAMETRIC UNCERTAINTIES. IF THE ENTRIES OF $\boldsymbol{c}_0$ ARE DETERMINED BY PASSIVE DEVICE MATCHING, THEN CASE 3 SUGGESTS THE UNCERTAINTIES THAT ONE MIGHT EXPECT IN PRACTICE

NTF OOBG of Modulator as a Function of Variable-Integrator Damping Given Three Cases of Parameteric Uncertainty				
m (mode)	$\{\Delta_i\}_{i=1}^7$	**Case 1** $\Delta c_{0i} = 2.5\%$ $\Delta a_{ij} = 2.5\%$	**Case 2** $\Delta c_{0i} = 5\%$ $\Delta a_{ij} = 5\%$	**Case 3** $\Delta c_{0i} = 1\%$ $\Delta a_{ij} = 20\%$
1	$\{0, 0, 0, 0, 0, 0, 0\}$	8.45	10.89	8.17
2	$\{0.2, 0.1, 0.1, 0.1, 0.025, 0.01, 0.005\}$	3.13	3.57	3.15
3	$\{0.2, 0.125, 0.125, 0.125, 0.08, 0.025, 0.007\}$	1.73	1.97	1.73
4	$\{0.4, 0.2, 0.2, 0.2, 0.1, 0.035, 0.02\}$	1.50	1.72	1.48
5	$\{0.5, 0.25, 0.25, 0.25, 0.15, 0.1, 0.03\}$	1.17	1.30	1.16
6	$\{1, 1, 1, 1, 1, 1, 1\}$	1.00	1.00	1.00

loop filter tuning. In an actual continuous-time modulator integrated circuit, the ability to tune the loop filter is essential.

We now illustrate the use of the VIDSR compensator for the on-line reduction of the NTF OOBG. Throughout this section, simulated data are obtained with the seventh-order single-bit modulator described by the (nominal) system parameters of Table III.

We recall that in the SR-compensated modulator, there are only two modes: $m = 0$ (basic operation) and $m = 1$ (soft-resetting). In the VIDSR scheme, the *detector* block incorporates a number of operational modes. A VIDSR mode is defined by two coordinates: 1) a set of values for $\Delta_i, i = 1, \ldots, n$, and 2) a corresponding noise transfer function (and OOBG). Each mode corresponds to a distinct value for the output of the *detector*, m. Table IV shows mode and Δ_i settings for the seventh-order single-bit modulator of Table III, corresponding to different uncertainty bounds on system parameters. A threshold of 10.0, based on the 1-norm, $||\boldsymbol{x}(t)|| = \sum_{i=1}^n |x_i(t)|$, is used to indicate the need for a transition to a subsequent mode.

Between transitions, the soft reset is maintained until the norm reduces to a lower threshold (in this case 1.0). The mode variable m increments with each threshold-crossing following a soft-reset. The *detector* block must estimate the norm in order to direct mode selection.

Fig. 13 illustrates the self-tuning capability of the VIDSR-compensated modulator. We simulate worst-case loop filter pa-

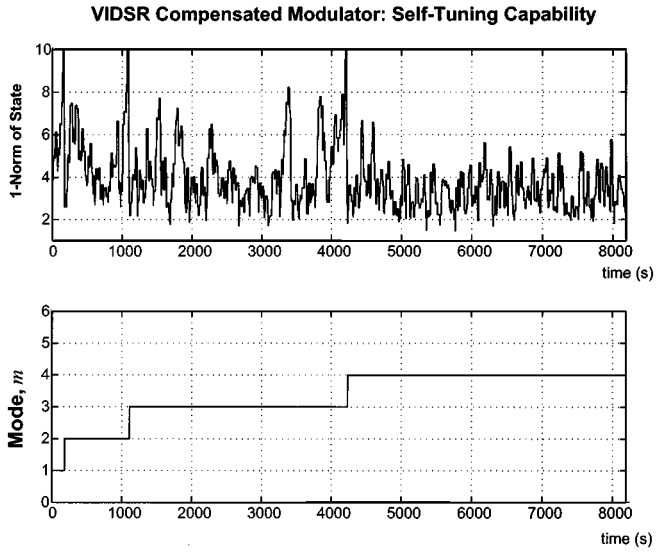

Fig. 13. Segment of simulation data for the VIDSR-compensated seventh-order modulator. Soft-resetting is deactivated with the rising edge of the mode transition signal. The modulator input signal, centred within the signal band (a sampling frequency of 2π-rad/s is applied), is $0.15\sin((\pi/128)t)$. Please note that the graph of the 1-norm has been median filtered to improve the quality of the plot.

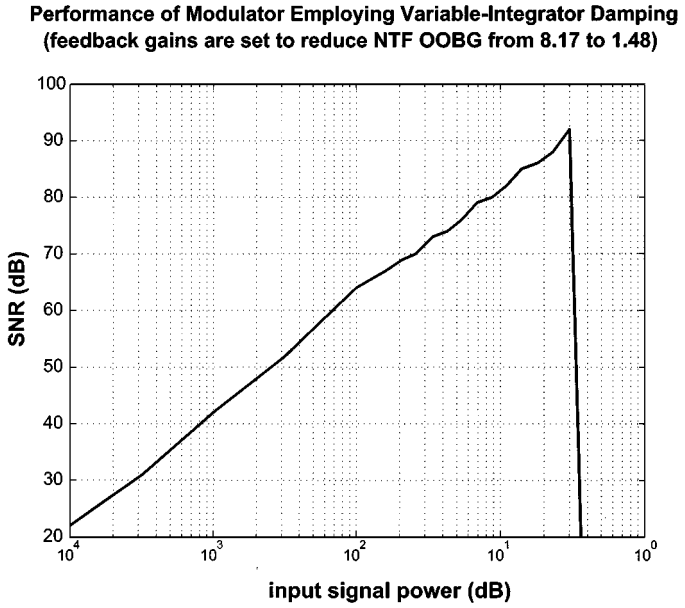

Fig. 14. Peak SNR of tuned VIDSR-compensated modulator. The NTF OOBG was tuned down from 8.17 to 1.48, as shown in Case 3 of Table IV, before the plot was generated.

rameters corresponding to $\Delta c_{0i} = 1\%$ and $\Delta a_{ij} = 20\%$, i.e., Case 3 of Table IV. To ensure a smooth transition between modes and to shrink states which may have surged, soft-resetting is applied immediately after a given threshold is exceeded. As shown in the figure, the system settles to a mode of $m = 4$, tuning itself to reduce the aggressiveness of the NTF and thus providing a method to compensate for uncertainty in loop filter parameters. Once tuned, the system achieves the SNR versus input signal power characteristic of Fig. 14, attaining a peak resolution of over 92 dB.

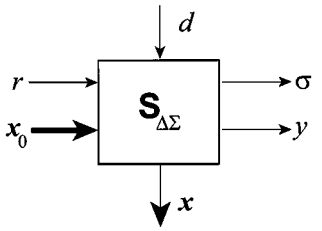

Fig. 15. Generalized delta–sigma modulator system with state vector $\boldsymbol{x}$, initial condition $\boldsymbol{x}_0$, scalar inputs r and d, and scalar outputs σ and y.

III. Formal Development

In this section, we prove the stabilizing effect of SR-compensation. We emphasize that an implicit assumption made in our development is that the sampling period of the system is zero. This infinite-sampling rate condition simplifies our formal development considerably, and is typical of stability proofs of variable-structure systems. In practice, finite sampling rates are generally acceptable as long as the sample period is much shorter than the fastest time constant associated with the plant. For oversampled systems, this is usually the case.

A. Preliminaries

We define the set $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ and make use of the following norms. Given $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$, we set

$$|f(\cdot)|_\infty := \sup_{t\geq 0} |f(t)|. \tag{6}$$

Given $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f}(\cdot) : \mathbb{R} \mapsto \mathbb{R}^n$, we define

$$\|\boldsymbol{x}\|_p := \begin{cases} \left[\sum_{i=1}^{n} |x_i|^p\right]^{\frac{1}{p}}, & p \in \{1,2\} \\ \max_i |x_i|, & p = \infty \end{cases} \tag{7}$$

and

$$\|\boldsymbol{f}(t)\|_p := \begin{cases} \left[\sum_{i=1}^{n} |f_i(t)|^p\right]^{\frac{1}{p}}, & p \in \{1,2\} \\ \sup_{t\geq 0} \|\boldsymbol{f}(t)\|_2, & p = \infty \end{cases}. \tag{8}$$

The variables x_i and $f_i(t)$ represent the i-th scalar components of $\boldsymbol{x}$ and $\boldsymbol{f}(t)$, respectively. The notation $\|\boldsymbol{x}\|$ denotes some norm on $\mathbb{R}^n$.

Definition 1 (Stability of $S_{\Delta\Sigma}$): The system of Fig. 15, namely, $S_{\Delta\Sigma}$, is M-stable, where $M \in [0,\infty)$, if $\|\boldsymbol{x}(\cdot)\|_\infty < \infty$, $|y(\cdot)|_\infty < \infty$ and $|\sigma(\cdot)|_\infty \leq M$ for any $\boldsymbol{x}_0, r$ and d satisfying $\|\boldsymbol{x}_0\|_2 \leq M_0$, $|r(\cdot)|_\infty < M_r$, and $|d(\cdot)|_\infty < M_d$, for some finite real positive constants M_0, M_r and M_d. □

B. Setup

We now define the delta–sigma modulator, under soft-reset compensation, as follows. The filter H is described as

$$H : \begin{cases} \dot{\boldsymbol{x}} = [\mathbf{A}_0 + \mathbf{K}(\boldsymbol{x})]\boldsymbol{x} + \boldsymbol{b}_0 r - \boldsymbol{b}_0 y \\ \sigma = \boldsymbol{c}_0 \boldsymbol{x} \\ y = Q(\sigma + d) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \in \mathbb{R}^n \end{cases} \tag{9}$$

in which $\boldsymbol{x} \in \mathbb{R}^n$ and $\sigma, r, d \in \mathbb{R}$. We assume that there exist finite real positive constants M_r and M_d such that $|r(\cdot)|_\infty < M_r$ and $|d(\cdot)|_\infty < M_d$. The state model parameters $\mathbf{A}_0 \in \mathbb{R}^{n\times n}$, $\mathbf{K}(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^{n\times n}$, $\boldsymbol{b}_0 \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{c}_0 \in \mathbb{R}^{1\times n}$ are given by

$$\mathbf{A}_0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{10}$$

$$\mathbf{K}(\boldsymbol{x}) = \operatorname{diag}(k_1(\boldsymbol{x}), \ldots, k_n(\boldsymbol{x})) \tag{11}$$

$$\boldsymbol{b}_0 = [0 \quad 0 \quad \cdots \quad 0 \quad 1]^T \quad \text{and} \tag{12}$$

$$\boldsymbol{c}_0 = [c_{01} \quad c_{02} \quad \cdots \quad c_{0(n-1)} \quad c_{0n}],$$
$$c_{0i} \in \mathbb{R}; \quad i = 1, \ldots, n, \tag{13}$$

where $k_i(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}, i = 1, \ldots, n$ and $\operatorname{diag}(\cdot, \ldots, \cdot)$ denotes a diagonal matrix (the $(1,1)$-element is given by the first argument, the $(2,2)$-element by the second argument, and so on). We set $Q(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ to

$$Q(\rho) = \begin{cases} +\delta, & \rho \geq T_\rho \\ -\delta, & \rho \leq -T_\rho \end{cases} \tag{14}$$

where $\delta > M_r, T_\rho \in (0, \delta)$. We require that $Q(\rho)$ be defined such that its magnitude over the set $\rho \in (-T_\rho, T_\rho)$ be less than δ, i.e., $|Q(\cdot)|_\infty = \delta$. The feedback elements are given by

$$k_i(\boldsymbol{x}) = \begin{cases} -\varepsilon_1 \operatorname{sgn}(c_{01}\sigma x_1), & i = 1 \\ -\frac{c_{0(i-1)}}{c_{0i}} - \varepsilon_i \operatorname{sgn}(c_{0i}\sigma x_i), & i = 2, \ldots, n \end{cases} \tag{15}$$

in which $\varepsilon_i > 0, i = 1, \ldots, n$. The sets $\boldsymbol{\Omega}_M$ and $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$, depicted in Fig. 8, are defined as

$$\boldsymbol{\Omega}_M = \{\boldsymbol{x} \in \mathbb{R}^n : |\sigma| \leq M\} \tag{16}$$

and

$$\boldsymbol{\Omega}_{\boldsymbol{\Delta}} = \{\boldsymbol{x} \in \mathbb{R}^n : |\sigma| \leq \boldsymbol{\Delta}\} \subset \boldsymbol{\Omega}_M \tag{17}$$

for $\boldsymbol{\Delta} = M_d + \max(T_\rho, (2\delta M_d)/(\delta - M_r))$ and some $M \in (0, \infty)$.

C. Assumptions

Stability is proved on the basis of the following conditions.

1) All zeros of the nominal loop filter, $H_0(s)$, obtained from H by setting to $\mathbf{K}(\cdot) = \mathbf{0}$, have negative real parts, i.e., H_0 is strictly minimum phase.
2) $c_{0n} > 0$.
3) $\boldsymbol{x}_0 \in \boldsymbol{\Omega}_M$.

Note that it can be shown from the first two assumptions that all coefficients of the numerator polynomial of $H_0(s)$ are greater than zero. The initial condition $\boldsymbol{x}_0$, in this context, refers to the point at which the SR compensator is activated. In this section, we are only concerned with the behavior of the modulator during a soft-reset.

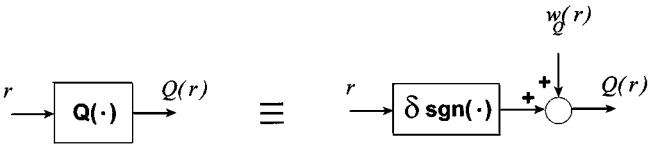

Fig. 16. Quantizer model.

Lemma 1 (Quantizer Model): The function $w_Q(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ defined as

$$w_Q(\rho) = Q(\rho) - \delta \operatorname{sgn}(\rho) \tag{18}$$

is 1) bounded and 2) equal to 0 for $|\rho| \geq T_\rho$. □

Proof: We note that $|w_Q(\rho)| \leq |Q(\cdot)|_\infty + \delta = 2\delta$ and that

$$w_Q(\rho) = Q(\rho) - \delta \operatorname{sgn}(\rho) = \begin{cases} \delta - \delta = 0, & \rho \geq T_\rho \\ -\delta + \delta = 0, & \rho \leq -T_\rho \end{cases}. \tag{19}$$

◇

This lemma allows us to treat any multi-level quantizer $Q(\cdot)$ as a single-bit quantizer, $\delta \operatorname{sgn}(\cdot)$, plus a bounded disturbance, $w_Q(\cdot)$. That is, $Q(\rho) = \delta \operatorname{sgn}(\rho) + w_Q(\rho)$, as shown in Fig. 16.

Lemma 2: Suppose that $\boldsymbol{x}_0 \in \mathbb{R}^n - \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. The modulator under soft-reset compensation cannot exhibit finite-escape-time trajectories, i.e., it cannot have solutions $\boldsymbol{x}(t)$ for which there exists a $t^* \in (0, \infty)$ such that $\lim_{t\to t^*-} \|\boldsymbol{x}(t)\|_2 = \infty$ in $\mathbb{R}^n - \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. □

Proof: We write the dynamics of the modulator (under soft-resetting) in $\mathbb{R}^n - \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ as

$$\dot{\boldsymbol{x}} = \mathbf{F}(t, \boldsymbol{x}) = [\mathbf{A}_0 + \mathbf{K}(\boldsymbol{x})]\boldsymbol{x} + \boldsymbol{b}_0 r - \boldsymbol{b}_0 y \tag{20}$$

and note that outside $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$, $|\sigma| > \boldsymbol{\Delta}$, which suggests that for $\sigma > \boldsymbol{\Delta}$

$$y = \delta \operatorname{sgn}(\sigma + d) + w_Q(\sigma + d). \tag{21}$$

Since

$$\boldsymbol{\Delta} \geq M_d + T_\rho \Rightarrow \delta \operatorname{sgn}(\sigma + d) = \delta \quad \text{and}$$
$$w_Q(\sigma + d) = 0 \tag{22}$$
$$y = \delta. \tag{23}$$

Similarly, for $\sigma < -\boldsymbol{\Delta}$, $y = -\delta$. We assume, without loss of generality, that $\sigma > \boldsymbol{\Delta}$ for the following discussion. Therefore,

$$\dot{\boldsymbol{x}} = \mathbf{F}(t, \boldsymbol{x}) = [\mathbf{A}_0 + \mathbf{K}(\boldsymbol{x})]\boldsymbol{x} + \boldsymbol{b}_0(r - \delta); \quad \sigma > \boldsymbol{\Delta}. \tag{24}$$

We may write (20) as the equivalent integral equation:

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_0^t \mathbf{F}[\tau, \boldsymbol{x}(\tau)]d\tau. \tag{25}$$

Note that since $\mathbf{F}(\cdot,\cdot)$ is continuous in its arguments, any $\boldsymbol{x}(\cdot)$ satisfying (25) must be continuous in t. We may write the i-th component of $\mathbf{F}$ as

$$F_i(t,\boldsymbol{x}) = \begin{cases} x_{i+1} + \alpha_i|x_i|, & i = 1,\ldots,n-1 \\ \alpha_i|x_i| + \underbrace{(r-\delta)}_{\leq 0}, & i = n \end{cases}$$
$$\Rightarrow |F_i(t,\boldsymbol{x})| \leq (1+|\alpha_i|)\|\boldsymbol{x}\|_2 \tag{26}$$

where $\alpha_i \in \mathbb{R}$, and x_i is the i-th element of $\boldsymbol{x}$. Thus,

$$\|\mathbf{F}(t,\boldsymbol{x})\|_2 \leq \sqrt{n}\max_i |F_i(t,\boldsymbol{x})| = \sqrt{n}(1+|\alpha_k|)\|\boldsymbol{x}\|_2 \tag{27}$$

for some $k \in \{1,\ldots,n\}$. Therefore,

$$\begin{aligned}\|\boldsymbol{x}(t)\|_2 &\leq \|\boldsymbol{x}_0\|_2 + \int_0^t \|\mathbf{F}[\tau,\boldsymbol{x}(\tau)]\|_2\, d\tau \\ &\leq \|\boldsymbol{x}_0\|_2 + \int_0^t [\sqrt{n}(1+|\alpha_k|)]\|\boldsymbol{x}(\tau)\|_2\, d\tau. \end{aligned}\tag{28}$$

In the region $\sigma > \boldsymbol{\Delta}$, $\mathbf{F}(t,\boldsymbol{x})$ is uniformly Lipschitz continuous; specifically, we can show that there exists an $L \in \mathbb{R}_+$ such that

$$\|\mathbf{F}(t,\boldsymbol{x}) - \mathbf{F}(t,\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x}-\boldsymbol{y}\|_2 \tag{29}$$

for all $\boldsymbol{x},\boldsymbol{y} \in \mathbf{D}_+ := \{\boldsymbol{p} \in \mathbb{R}^n : \boldsymbol{c_0 p} > \boldsymbol{\Delta}\}$, for all $t \geq 0$.

Suppose that there exists a $t^* < \infty$ such that

$$\lim_{t\to t^*-} \|\boldsymbol{x}(t)\|_2 = \infty \tag{30}$$

where $\boldsymbol{x}(t) \in \mathbf{D}_+$ for all $t \in [0,t^*)$. From our Lipschitz condition, we may infer the existence of a solution $\boldsymbol{x}(t)$ on $[0,t^*)$. Since the solution must be continuous on $[0,t^*)$, we can apply Gronwall's Lemma [15] in view of (28) which gives us a bound on the solution,

$$\|\boldsymbol{x}(t)\|_2 \leq \|\boldsymbol{x}_0\|_2 \exp\{[\sqrt{n}(1+|\alpha_k|)]t\} \qquad \forall t \in [0,t^*). \tag{31}$$

Therefore,

$$\lim_{t\to t^*} \|\boldsymbol{x}(t)\|_2 \leq \|\boldsymbol{x}_0\|_2 \exp\{[\sqrt{n}(1+|\alpha_k|)]t^*\} < \infty \tag{32}$$

which contradicts (30), since the solution must have no jump discontinuities in view of (25) and the continuity of $\mathbf{F}(\cdot,\cdot)$. Therefore there can be no finite-escape-time trajectory in $\mathbb{R}^n - \boldsymbol{\Omega}_\Delta$. ◇

Theorem 1: The modulator under soft-reset compensation is M-stable. □

Proof: We define

$$V(\sigma) = \frac{1}{2}\sigma^2 \tag{33}$$

and differentiate with respect to time

$$\begin{aligned}\dot{V} = \frac{dV}{d\sigma}\dot{\sigma} &= \sigma \boldsymbol{c}_0\{[\mathbf{A}_0 + \mathbf{K}(\boldsymbol{x})]\boldsymbol{x} + \boldsymbol{b}_0 r \\ &\quad - \boldsymbol{b}_0\delta\,\mathrm{sgn}(\sigma+d) - \boldsymbol{b}_0 w_Q(\sigma+d)\} \\ &= c_{01}k_1(\boldsymbol{x})\sigma x_1 + \left\{\sum_{i=2}^{n} c_{0i}\left[\frac{c_{0(i-1)}}{c_{0i}} + k_i(\boldsymbol{x})\right]\sigma x_i\right\} \\ &\quad + c_{0n}[r - \delta\,\mathrm{sgn}(\sigma+d) - w_Q(\sigma+d)]\sigma.\end{aligned}$$

We now substitute for k_i using (15) to obtain

$$\dot{V} = -\sum_{i=1}^{n}\varepsilon_i|c_{0i}\sigma x_i| + c_{0n}[r - \delta\,\mathrm{sgn}(\sigma+d) - w_Q(\sigma+d)]\sigma. \tag{34}$$

Continuing

$$\begin{aligned}\dot{V} &= -\sum_{i=1}^{n}\varepsilon_i|c_{0i}\sigma x_i| + c_{0n}r(\sigma+d) \\ &\quad - c_{0n}\delta\,\mathrm{sgn}(\sigma+d)(\sigma+d) - c_{0n}w_Q(\sigma+d)\sigma \\ &\quad - c_{0n}rd + c_{0n}\delta\,\mathrm{sgn}(\sigma+d)d \\ &< -\sum_{i=1}^{n}\varepsilon_i|c_{0i}\sigma x_i| + c_{0n}|r(\cdot)|_\infty|\sigma+d| \\ &\quad - c_{0n}\delta\,\mathrm{sgn}(\sigma+d)(\sigma+d) + c_{0n}|w_Q(\sigma+d)||\sigma| \\ &\quad + c_{0n}[|r(\cdot)|_\infty + \delta]|d(\cdot)|_\infty \\ &= -\sum_{i=1}^{n}\varepsilon_i|c_{0i}\sigma x_i| + c_{0n}[|r(\cdot)|_\infty - \delta]|\sigma+d| \\ &\quad + c_{0n}|w_Q(\sigma+d)||\sigma| + c_{0n}[|r(\cdot)|_\infty + \delta]|d(\cdot)|_\infty.\end{aligned}$$

We write $\dot{V} < P_{\dot{V}} + Q_{\dot{V}}$ by setting

$$P_{\dot{V}} := \left[-\sum_{i=1}^{n}\varepsilon_i|c_{0i}x_i| + c_{0n}|w_Q(\sigma+d)|\right]|\sigma|.$$

and

$$Q_{\dot{V}} := c_{0n}[|r(\cdot)|_\infty - \delta]|\sigma+d| + c_{0n}[|r(\cdot)|_\infty + \delta]|d(\cdot)|_\infty. \tag{35}$$

We note [see (36) at the bottom of the page] and that $w_Q(\sigma+d) = 0$ for $|\sigma+d| \geq T_\rho$.

Since

$$|\sigma+d| \geq |\sigma| - |d| \geq |\sigma| - |d(\cdot)|_\infty \geq |\sigma| - M_d \tag{37}$$

$$P_{\dot{V}} \leq \left[\underbrace{-(\min_i \varepsilon_i)(\min_i |c_{0i}|)\left(\sum_{i=1}^{n}|x_i|\right)}_{\leq 0} + c_{0n}|w_Q(\sigma+d)|\right]|\sigma| \tag{36}$$

then if $|\sigma| - M_d \geq T_\rho$, equivalently, if $|\sigma| \geq M_d + T_\rho$, then $P_{\dot{V}} < 0$. We also have

$$\begin{aligned} Q_{\dot{V}} &\leq c_{0n}(M_r - \delta)|\sigma + d| + c_{0n}(M_r + \delta)M_d \\ &\leq c_{0n}(M_r - \delta)|\sigma + d| + 2c_{0n}\delta M_d \\ &= c_{0n} \underbrace{((M_r - \delta)|\sigma + d| + 2\delta M_d)}_{\text{we want this} < 0}. \end{aligned}$$

To ensure that $Q_{\dot{V}}$ is less than zero, we set

$$|\sigma + d| > \frac{2\delta M_d}{(\delta - M_r)}. \tag{38}$$

Since $|\sigma+d| \geq |\sigma|-M_d$, we note that $|\sigma| > [(2\delta/(\delta - M_r))+ 1]\, M_d \Rightarrow Q_{\dot{V}} < 0$. Therefore, since $\dot{V} < P_{\dot{V}} + Q_{\dot{V}}$, we have that $\dot{V} < 0$ if

$$|\sigma| \geq \boldsymbol{\Delta} = M_d + \max\left\{T_\rho, \frac{2\delta M_d}{(\delta - M_r)}\right\}. \tag{39}$$

By Lemma 2, $\boldsymbol{x}(t)$ cannot escape to infinity in finite time in $\mathbb{R}^n - \boldsymbol{\Omega}_\Delta$. Given that $\boldsymbol{x}$ is initially in the set $\mathbb{R}^n - \boldsymbol{\Omega}_\Delta$, we can determine a time $T_{\boldsymbol{\Delta}} > 0$ for which $\boldsymbol{x}(t) \in \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ for all $t > T_{\boldsymbol{\Delta}}$, as follows. We have shown that

$$Q_{\dot{V}} < 0, \quad |\sigma| \geq \boldsymbol{\Delta}. \tag{40}$$

Noting that

$$P_{\dot{V}} = -\left[\sum_{i=1}^{n} \varepsilon_i |c_{0i} x_i|\right] |\sigma|, \quad |\sigma| \geq \boldsymbol{\Delta} \tag{41}$$

$$\leq -\left[\left(\min_i \varepsilon_i\right)\left(\min_i |c_{0i}|\right) \|\boldsymbol{x}\|_1\right] |\sigma| \tag{42}$$

and introducing

$$\kappa := \min_{\boldsymbol{x} \in \mathbb{R}^n - \boldsymbol{\Omega}_\Delta} \left[\left(\min_i \varepsilon_i\right)\left(\min_i |c_{0i}|\right) \|\boldsymbol{x}\|_1\right] > 0 \tag{43}$$

we have

$$P_{\dot{V}} = -\kappa|\sigma|, \quad |\sigma| \geq \boldsymbol{\Delta} \tag{44}$$

$$\leq -\kappa \boldsymbol{\Delta}. \tag{45}$$

Since $\dot{V} < P_{\dot{V}} + Q_{\dot{V}}$

$$\dot{V} < -\kappa \boldsymbol{\Delta}, \quad |\sigma| \geq \boldsymbol{\Delta}. \tag{46}$$

Integrating both sides of this expression with respect to time on the interval $(0, T]$ gives (with a slight abuse of notation)

$$V(T) < V(0) - \kappa \Delta T, \quad |\sigma| \geq \boldsymbol{\Delta}. \tag{47}$$

Setting $V(T) = (1/2)\boldsymbol{\Delta}^2$ and solving for T reveals that $\boldsymbol{x}(t)$ must intersect $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ before a time

$$T_{\boldsymbol{\Delta}} := \frac{V(0) - \frac{1}{2}\boldsymbol{\Delta}^2}{\kappa \boldsymbol{\Delta}} > 0. \tag{48}$$

That is, $\boldsymbol{x}(t) \in \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ for all $t \geq T_{\boldsymbol{\Delta}}$. Once $\boldsymbol{x}(t) \in \boldsymbol{\Omega}_{\boldsymbol{\Delta}}$, $\boldsymbol{x}(t)$ must remain within $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$ for all subsequent time since $\dot{V} < 0$ for all $\boldsymbol{x} \in \mathbb{R}^n - \boldsymbol{\Omega}_\Delta$.

We now consider the evolution of $\boldsymbol{x}(t)$ within $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. We define T_e as the time at which $\boldsymbol{x}(t)$ first intersects $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. Therefore $T_e \in [0, T_{\boldsymbol{\Delta}}]$. Since $\dot{V} < 0$ outside $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$, $\boldsymbol{x}(t)$ cannot leave $\boldsymbol{\Omega}_{\boldsymbol{\Delta}}$. To show that $\boldsymbol{x}(t)$ is bounded, we will show that $\|\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t)\|_2$ is bounded, where $\bar{\boldsymbol{x}}(t)$ denotes the sliding-mode trajectory of the stable system

$$\boldsymbol{\Sigma}_{\text{sm}} : \begin{cases} \dot{\bar{\boldsymbol{x}}} = [\mathbf{I} - \boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\boldsymbol{c}_0]\mathbf{A}_0\bar{\boldsymbol{x}} \\ \boldsymbol{c}_0\bar{\boldsymbol{x}} = 0 \end{cases} \tag{49}$$

which we initialize to $\bar{\boldsymbol{x}}(T_e) = \boldsymbol{x}(T_e) - ((\boldsymbol{\Delta}/\|\boldsymbol{c}_0\|_2)\boldsymbol{c}_0^T)$. The trajectories of (49) and the modulator under soft-reset compensation are illustrated in Fig. 8 for the second-order case. For details concerning our use of the *method of equivalent control*, please see Appendix B. The system $\boldsymbol{\Sigma}_{\text{sm}}$ is exponentially stable because H_0 is strictly minimum phase; a proof of this is provided in Appendix C.

The following is adapted from Utkin ([6]). From the method of equivalent control, we can write the dynamics of our modulator in (9) under soft-reset compensation as

$$\boldsymbol{\Sigma}_{\text{sr}} : \begin{cases} \dot{\boldsymbol{x}} = [\mathbf{I} - \boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\boldsymbol{c}_0]\mathbf{A}_0\boldsymbol{x} + \boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\dot{\sigma} \\ \sigma = \boldsymbol{c}_0\boldsymbol{x} \end{cases} \tag{50}$$

for $t \geq T_e$. We let $\boldsymbol{\Psi}(\,\cdot\,) : \mathbb{R} \mapsto \mathbb{R}^{n \times n}$ denote the state transition matrix associated with (49). Thus the solution to (50) may be written as

$$\boldsymbol{x}(t) = \boldsymbol{\Psi}(t - T_e)\boldsymbol{x}(T_e) + \int_{T_e}^{t} \boldsymbol{\Psi}(t - \tau)\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\dot{\sigma}(\tau)\, d\tau, \quad t \in (T_e, \infty). \tag{51}$$

Re-expressing the integral of (51) using integration-by-parts, we obtain

$$\begin{aligned} \boldsymbol{x}(t) = {} & \boldsymbol{\Psi}(t - T_e)\boldsymbol{x}(T_e) + \boldsymbol{\Psi}(0)\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\sigma(t) \\ & - \boldsymbol{\Psi}(t - T_e)\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\sigma(T_e) \\ & - \int_{T_e}^{t} \dot{\boldsymbol{\Psi}}(t - \tau)\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\sigma(\tau)\, d\tau, \quad t \in (T_e, \infty). \end{aligned} \tag{52}$$

The solution to (49) is written simply as

$$\bar{\boldsymbol{x}}(t) = \boldsymbol{\Psi}(t - T_e)\bar{\boldsymbol{x}}(T_e), \quad t \in (T_e, \infty). \tag{53}$$

Therefore, from (52) and (53) we determine the bound on $\|\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t)\|_2$ for $t \in (T_e, \infty)$ as

$$\begin{aligned} & \|\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t)\|_2 \\ & \quad \leq \boldsymbol{\Delta}\left\{(1 + \|\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\|_2)\|\boldsymbol{\Psi}(t - T_e)\|_{2,i}\right. \\ & \qquad \left. + \left(1 + \int_{T_e}^{t} \|\dot{\boldsymbol{\Psi}}(t - \tau)\|_{2,i}\, d\tau\right)\|\boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\|_2\right\} \end{aligned} \tag{54}$$

where $\|\,\cdot\,\|_{2,i}$ denotes the induced Euclidean norm. Since (49) is stable, $\|\boldsymbol{\Psi}(t)\|_{2,i}$ and $\int_{T_e}^{t} \|\dot{\boldsymbol{\Psi}}(\tau)\|_{2,i}\, d\tau$ are bounded. Thus, we have that

$$\|\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t)\|_2 \leq N\boldsymbol{\Delta} \tag{55}$$

for some $N \in (0, \infty)$, for all $t \geq T_e$. Therefore, $\|\boldsymbol{x}(t)\|_2$ is bounded. ◇

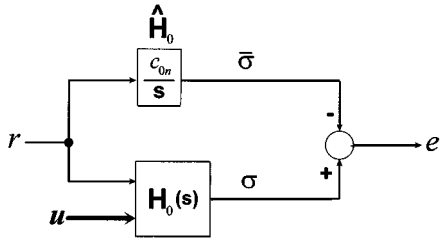

Fig. 17. Model-matching set-up to find the effective NTF. We seek the input $\boldsymbol{u}$ which renders the nominal loop filter $H_0(s)$ equivalent to the first-order system, $\hat{H}_0(s)$. Note that $\boldsymbol{u}$ is a vector-input.

Remark: We note that if $\boldsymbol{\Delta}$ is sufficiently small, (55) suggests that $\boldsymbol{x}(t)$ enters and remains within a given neighborhood of the origin. This confirms that soft-resetting can provide a state-shrinking effect.

IV. Conclusion

We have presented two compensation architectures for continuous-time delta–sigma modulators employing loop filters of any order. These methods are based on variable-structure control techniques and offer 1) soft-resetting as an alternative to conventional resetting and 2) measures to counter parametric uncertainty. Although an infinite sampling rate condition is imposed, the power of our approach is that it accommodates arbitrary inputs (with bounded peak magnitudes), dithering and multi-level quantization. These compensators are intended for high-order modulators employing coarse quantizers and may be helpful in the design of wideband transceiver systems.

Appendix A
Effective NTF Under Soft-Reset Compensation

We begin with the set-up of Fig. 17. Our task is to find $\boldsymbol{u}$ such that the nth-order nominal loop filter H_0 is equivalent to $\hat{H}_0$. We will prove that the NTF under soft resetting, i.e., the *effective* NTF, given by $1/(1+H_0(s))$, is equal to $s/(s+c_{0n})$. We make the assumption that the switching offsets ε_i appearing in (15) are all zero; in practice, the ε_i are small, and the switching feedbacks k_i vary slightly about fixed "average" values. We introduce the state-space systems

$$H_0 : \begin{cases} \dot{\boldsymbol{x}} = \mathbf{A}_0\boldsymbol{x} + [\boldsymbol{b}_0 \quad \mathbf{I}_n]\begin{bmatrix} r \\ \boldsymbol{u} \end{bmatrix} \\ \sigma = \boldsymbol{c}_0\boldsymbol{x} \end{cases} \tag{56}$$

and

$$\hat{H}_0 : \begin{cases} \dot{\hat{x}} = c_{0n} r \\ \hat{\sigma} = \hat{x} \end{cases} \tag{57}$$

where $\boldsymbol{u} = \overline{\mathbf{K}}\boldsymbol{x}$ and $\overline{\mathbf{K}} = \operatorname{diag}(\bar{k}_1, \ldots, \bar{k}_n)$. All other parameters are defined as in Section 3. We assume that $\hat{H}_0(s)$ and $H_0(s)$ are initialized such that $\sigma = \hat{\sigma} = 0$. We define the output error as

$$e = \sigma - \hat{\sigma}. \tag{58}$$

Now differentiate the error with respect to time

$$\dot{e} = \dot{\sigma} - \dot{\hat{\sigma}} \tag{59}$$

$$= \boldsymbol{c}_0(\mathbf{A}_0\boldsymbol{x} + \boldsymbol{b}_0 r + \mathbf{I}_n\boldsymbol{u}) - c_{0n} r \tag{60}$$

and make the substitution $\boldsymbol{u} = \overline{\mathbf{K}}\boldsymbol{x}$ to obtain

$$\dot{e} = \boldsymbol{c}_0(\mathbf{A}_0 + \overline{\mathbf{K}})\boldsymbol{x} + \boldsymbol{c}_0\boldsymbol{b}_0 r - c_{0n} r = \boldsymbol{c}_0(\mathbf{A}_0 + \overline{\mathbf{K}})\boldsymbol{x}. \tag{61}$$

But we note that

$$\boldsymbol{c}_0(\mathbf{A}_0 + \overline{\mathbf{K}}) = [c_{01}\bar{k}_1 \quad (c_{01} + c_{02}\bar{k}_2) \quad \ldots \quad (c_{0n-1} + c_{0n}\bar{k}_n)]. \tag{62}$$

If we substitute for each $\bar{k}_i$ the corresponding k_i given in (15) with $\varepsilon_i = 0$ for all i, we obtain

$$\dot{e} = 0 \Rightarrow e(t) = 0 \quad \forall t \geq 0. \tag{63}$$

Therefore H_0 is equivalent to $\hat{H}_0$ and we obtain the effective noise transfer function

$$N_G(s) := \frac{1}{1 + H_0(s)} = \frac{1}{1 + \frac{c_{0n}}{s}} = \frac{s}{s + c_{0n}}. \tag{64}$$

◇

Appendix B
Use of the Method of Equivalent Control

Variable-structure theory conventionally requires that the system under consideration have an equal number of outputs (switching surfaces) and control inputs. The proposed modulator architecture is multi-input, single-output. In this section we present the technical details of how variable-structure theory can be made to accommodate our design, and why the so-called "sliding-mode" solution ([5]) $\bar{\boldsymbol{x}}(t)$ is indeed given by the linear system of (49).

We can write the dynamics of a delta–sigma modulator in the expanded form

$$S_e : \begin{cases} \dot{\bar{\boldsymbol{x}}} = \mathbf{A}_0\bar{\boldsymbol{x}} + \mathbf{B}_0\boldsymbol{u} + \boldsymbol{b}_0 r \\ \bar{\boldsymbol{\Sigma}} = \mathbf{C}_0\bar{\boldsymbol{x}} \end{cases} \tag{65}$$

where $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$ and $\mathbf{C}_0 \in \mathbb{R}^{n\times n}$. Assuming that the system is undergoing a soft-reset, we have

$$\mathbf{B}_0 = \mathbf{I}, \tag{66}$$

$$\boldsymbol{u} = \mathbf{K}(\bar{\boldsymbol{x}})\bar{\boldsymbol{x}} - \boldsymbol{b}_0 y \quad \text{and} \tag{67}$$

$$\mathbf{C}_0 = \begin{bmatrix} \boldsymbol{c}_0 \\ \boldsymbol{c}_0 \\ \vdots \\ \boldsymbol{c}_0 \end{bmatrix}. \tag{68}$$

Our augmented system S_e includes $(n-1)$ additional "dummy" outputs. According to Utkin ([6]), a sliding mode solution exists on $\bar{\sigma} := \boldsymbol{c}_0\bar{\boldsymbol{x}} = 0$ if $\bar{\sigma}\dot{\bar{\sigma}} < 0$, for all $\bar{\sigma} \neq 0$. From (65) we obtain (also noting that the sliding-mode solution corresponds to setting $y = \delta\operatorname{sgn}(\bar{\sigma})$ and $d = 0$)

$$\bar{\sigma}\dot{\bar{\sigma}} = c_{01}k_1(\bar{\boldsymbol{x}})\bar{\sigma}x_1 + \left\{\sum_{i=2}^{n} c_{0i}\left[\frac{c_{0(i-1)}}{c_{0i}} + k_i(\bar{\boldsymbol{x}})\right]\bar{\sigma}x_i\right\} + c_{0n}[r - \delta\operatorname{sgn}(\bar{\sigma})]\bar{\sigma} \tag{69}$$

$$= -|c_{01}||\varepsilon_1||\bar{\sigma}||x_1| - \sum_{i=2}^{n}|c_{0i}|\varepsilon_i|\bar{\sigma}||x_i| + c_{0n}[r - \delta\operatorname{sgn}(\bar{\sigma})]\bar{\sigma} \tag{70}$$

$$< -\sum_{i=1}^{n} |c_{0i}|\varepsilon_i|\bar{\sigma}||x_i| + c_{0n}|r(\cdot)|_\infty|\bar{\sigma}| - c_{0n}\delta|\bar{\sigma}| \quad (71)$$

$$= -\sum_{i=1}^{n} |c_{0i}|\varepsilon_i|\bar{\sigma}||x_i| + c_{0n}|\bar{\sigma}| \times [|r(\cdot)|_\infty - \delta] < 0, \quad \bar{\sigma} \neq 0. \quad (72)$$

Thus a sliding mode exists on $\bar{\sigma} = 0$. From the state model of S_e, we have that $\mathrm{rank}(\mathbf{C}_0) = 1$, $\mathrm{rank}(\mathbf{B}_0) = n$ and $\mathrm{rank}(\mathbf{C}_0\mathbf{B}_0) = 1$. Under these circumstances, we may, as indicated by Utkin in [6], arbitrarily assign $(n-1)$ components of the control vector $\boldsymbol{u}$ so that the solution obtained from setting $\dot{\bar{\sigma}} = \bar{\sigma} = 0$ can be found using the *method of equivalent control*. Thus, for the purposes of determining the sliding mode solution, S_e is identical to the single-input single-output system

$$\dot{\bar{\boldsymbol{x}}} = \mathbf{A}_0\bar{\boldsymbol{x}} + \boldsymbol{b}_0(r + u - y)$$
$$\bar{\sigma} = \boldsymbol{c}_0\bar{\boldsymbol{x}} \quad (73)$$

if we set the first $(n-1)$ components of $\boldsymbol{u}$ to zero. Here, the scalar u denotes our single control input (the n-th component of $\boldsymbol{u}$) to be used as the "equivalent control," as shown below. Following the method of equivalent control, we obtain

$$\dot{\bar{\sigma}} = 0 \quad (74)$$
$$\Rightarrow \boldsymbol{c}_0[\mathbf{A}_0\bar{\boldsymbol{x}} + \boldsymbol{b}_0(r + u - y)] = 0 \quad (75)$$
$$\Rightarrow u = (\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\boldsymbol{c}_0\mathbf{A}_0\bar{\boldsymbol{x}} + y - r. \quad (76)$$

Substituting this expression for u into (73) yields the linear time-invariant sliding mode dynamics of (49) when combined with the constraint that $\bar{\sigma} = 0$. If H_0 is strictly minimum phase, then $||\bar{\boldsymbol{x}}(t)||$ is bounded and goes to zero exponentially as $t \to \infty$. A proof of this statement is given in Appendix C.

Appendix C
Stability of Sliding Mode From Strict Minimum Phase Assumption on H_0

In this section we show that the dynamics of (49) are exponentially stable if H_0 is strictly minimum phase. We first note that

$$[\mathbf{I} - \boldsymbol{b}_0(\boldsymbol{c}_0\boldsymbol{b}_0)^{-1}\boldsymbol{c}_0]\mathbf{A}_0 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & -\frac{c_{01}}{c_{0n}} & -\frac{c_{02}}{c_{0n}} & \dots & -\frac{c_{0(n-1)}}{c_{0n}} \end{bmatrix}. \quad (77)$$

Therefore, the sliding mode dynamics are actually of order $n-1$. The first state variable, x_1, can be expressed as the linear combination

$$x_1 = -\frac{c_{02}}{c_{01}}x_2 - \frac{c_{03}}{c_{01}}x_3 - \cdots - \frac{c_{0n}}{c_{01}}x_n. \quad (78)$$

From inspection of (77), the dynamic system formed by states $x_2, \dots, x_n$ has poles given by the roots of the polynomial

$$p_1(z) = z^{n-1} + \frac{c_{0(n-1)}}{c_{0n}}z^{n-2} + \frac{c_{0(n-2)}}{c_{0n}}z^{n-3} + \cdots + \frac{c_{02}}{c_{0n}}z + \frac{c_{01}}{c_{0n}}. \quad (79)$$

Now, the zeros of H_0 are given by the roots of the numerator polynomial of its transfer function (which is readily obtained from its state model $(\mathbf{A}_0, \boldsymbol{b}_0, \boldsymbol{c}_0, 0)$). This polynomial is

$$p_2(\lambda) = c_{0n}\lambda^{n-1} + c_{0(n-1)}\lambda^{n-2} + \cdots + c_{01}. \quad (80)$$

Since $p_1(z)$ and $p_2(\lambda)$ have the same roots, the zeros of H_0 are the eigenvalues of the sliding mode dynamics. Thus if H_0 is strictly minimum phase, the sliding mode dynamics are exponentially stable.

Acknowledgment

The authors would like to thank Professor B. A. Francis at the University of Toronto for his helpful comments regarding the formal development of this paper. One of the authors (T. Z.) would also like to thank the reviewers and Professor M. J. Ogorzalek for their helpful comments and enthusiasm.

References

[1] J. F. Jensen, G. Raghavan, A. E. Cosand, and R. H. Walden, "A 3.2-GHz second-order delta-sigma modulator implemented in InP HBT technology," *IEEE J. Solid-State Circuits*, vol. 30, pp. 1119–1127, Oct. 1995.

[2] O. Shoaei, "Continuous-time Delta-Sigma A/D converters for high speed applications," Ph.D. dissertation, Carleton University, 1995.

[3] J. A. Cherry and W. M. Snelgrove, "Excess loop delay in continuous-time delta-sigma modulators," *IEEE Trans. Circuits Syst. II*, vol. 44, pp. 376–389, Apr. 1999.

[4] M. Erbar, M. Rieger, and H. Schemmann, "A 1.28-GHz sigma-delta modulator for video A/D conversion," in *Proc. IEEE Int. Conf. Consumer Electron.*, 1996, pp. 78–79.

[5] R. A. DeCarlo, S. H. Zak, and G. P. Matthews, "Variable structure control of nonlinear multivariable systems: A tutorial," *Proc. IEEE*, vol. 76, pp. 212–232, Mar. 1988.

[6] V. I. Utkin, *Sliding Modes in Control and Optimization.* New York: Springer-Verlag, 1992.

[7] H. Sira-Ramirez and M. Rios Bolivar, "Sliding mode control of dc-to-dc power converters via extended linearization," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 652–661, Oct. 1994.

[8] H. Sira Ramirez, "Switched control of bilinear converters via pseudolinearization," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 858–865, June 1989.

[9] C. Wolff, J. G. Kenney, and L. R. Carley, "CAD for the analysis and design of $\Delta\Sigma$ converters," in *Delta-Sigma Data Converters: Theory, Design and Simulation*, S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds. New York: IEEE Press, 1997, pp. 447–467.

[10] R. Schreier, M. V. Goodson, and B. Zhang, "An algorithm for computing convex positively invariant sets for delta-sigma modulators," *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 38–44, Jan. 1997.

[11] R. W. Adams and R. Schreier, "Stability theory for $\Delta\Sigma$ modulators," in *Delta-Sigma Data Converters: Theory, Design and Simulation*, S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds. New York: IEEE, 1997, pp. 141–164.

[12] S. M. Moussavi and B. H. Leung, "High-order single-stage single-bit oversampling A/D converter stabilized with local feedback loops," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 19–25, Jan. 1994.

[13] L. J. Breems, E. J. van der Zwan, and J. H. Huijsing, "A 1.8-mW CMOS $\Sigma\Delta$ modulator with integrated mixer for A/D conversion of IF signals," *IEEE J. Solid-State Circuits*, vol. 35, pp. 468–475, Apr. 2000.

[14] W. Redman-White and A. M. Durham, "Integrated fourth-order $\Sigma\Delta$ convertor with stable self-tuning continuous-time noise shaper," *Proc. Inst. Electr. Eng.* , vol. 141, no. 3, pp. 145–150, 1994.

[15] H. K. Khalil, *Nonlinear Systems*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.

Takis Zourntos (S'91–M'00) received the B.A.Sc. degree in engineering science, with an electrical option and the M.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, Canada. He is currently working toward the Ph.D. degree in electrical engineering, at the University of Toronto, combining control and signal processing theory with integrated circuit engineering to develop novel compensators for delta-sigma modulators.

He is a co-founder of Protolinx Corporation (incorporated September 2000), a start-up comany producing ultrafast wireless network products. His homepage is located at http://www.eecg.toronto.edu/~takis.

David A. Johns (S'81-M'89-SM'94-F'01) received the B.A.Sc., M.A.Sc., and Ph.D. degrees from the University of Toronto, Toronto, Canada, in 1980, 1983, and 1989, respectively.

In 1988, he joined the University of Toronto where he is currently a full professor. He has ongoing research programs in the general area of analog integrated circuits with particular emphasis on circuits and systems for digital communications. Together with academic experience, he has four years of semiconductor industrial experience during 1980, 1983-1985, and 1995 and is co-founder of a microelectronics company called Snowbush. His research work has resulted in more than 40 publications.

Dr. Johns is the recipient of the 1999 IEEE Darlington Award. He is the co-author of a textbook entitled "Analog Integrated Circuit Design" (NewYork: Wiley, 1997). He served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995 and for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from 1995 to 1997. His homepage is located at http://www.eecg.toronto.edu/~johns.