Oversampled Encoding without Delta-Sigma Modulation: A Novel Alternative based on Nonlinear Control

Takis Zourntos

Abstract— **We describe a new oversampled encoding scheme that explicitly minimizes quantization error through the use of a nonlinear controller. The approach is theoretically stable, suitable for mixed-signal integrated circuit implementation and effectively suppresses spurious tones. Theory and simulation results are presented.**

I. Introduction

Oversampled data conversion has emerged as a major enabling technology in applications such as digital audio and mobile digital communications [1] [2]. Realized using integrated circuits in a variety of semiconductor technologies, oversampled analog-to-digital (A/D) and digital-to-analog (D/A) converters are critical signal processing elements occurring in a wide range of electronics devices in professional and consumer markets.

The delta-sigma modulator, a nonlinear feedback system shown in Figure 1, is the basis for all oversampled encoders in current use and may be configured as either an A/D or D/A converter. High converter accuracy/resolution (exceeding 15 bits) can be achieved in an averaged-sense, since, because of oversampling, a large number of output words are used to represent a single sample of the input signal.

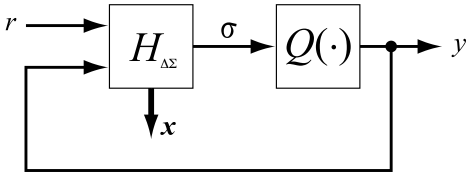

Fig. 1. The delta-sigma modulator. The input signal, r, is represented in a finite-extent frequency band with the coarsely quantized output signal, y. The state vector of the linear time-invariant loop filter, $H_{\triangle\Sigma}$, is denoted by $\boldsymbol{x}$ (vector signals are indicated with bold lines in all figures). The quantizer, $Q(\cdot)$, is a clocked device which operates at the sampling rate of the system.

The ultimate accuracy of the delta-sigma modulator is limited by stability. As the aggressiveness of the design increases, the margin of stability diminishes rapidly [3]. The performance forfeited to maintain desirable modulator behavior (i.e., the infrequent onset of large states) is substantial. Known, sufficient conditions for stability are viewed as too conservative by practitioners and are usually violated to extract greater peak performance at the expense of the occurrence of (occasional to frequent) glitches [4]. Stabilization methods for delta-sigma modulators such as those proposed in [5] or [6] are not considered here since the focus of this paper is to present a high-performance alternative to delta-sigma modulation.

This paper proposes a novel approach to reliable oversampled encoding based on nonlinear control theory. This work is significant because oversampled encoding is cost-effectively achieved without the use of delta-sigma modulation and stability can be guaranteed in a general sense.

II. Problem Formulation

We begin by formulating the problem of data conversion as a tracking-control problem. Essentially, our task is to match a held, coarsely quantized signal to an input (reference) signal within a specified bandwidth. In other words, given the output of a quantizer, y, and input signal, r, we would like to minimize the magnitude of the error e_o defined as

$$e_o = H(r - y) \tag{1}$$

where we assume that H (not to be confused with $H_{\triangle\Sigma}$ described above) denotes a continuous-time linear time-invariant filter of order n, $0 < n < \infty$. We assume that the frequency response of H restricts the comparison between r and y to a specified *signal band*, as depicted in Figure 2.

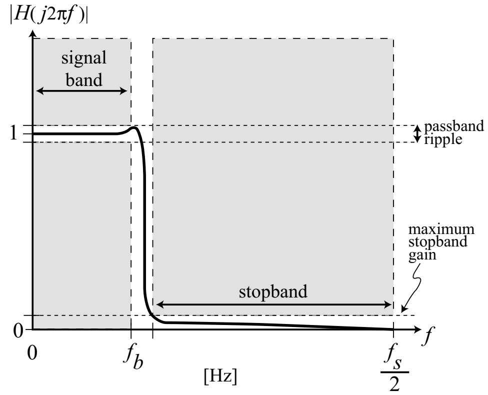

Fig. 2. H provides approximately unity-gain within the signal band and roughly zero gain outside. Note that H may also exhibit a bandpass characteristic instead of the lowpass response shown here.

This definition of tracking error naturally leads to the tracking control problem shown in Figure 3, where the "mystery block" denoted by a question mark represents a desired controller. This system is a novel representation of a signal encoding problem, cast in control-theoretic terms. We shall propose a controller design below to stabilize the

T. Zourntos is with Department of Electrical Engineering, Texas A & M University, College Station, Texas, USA 77843-3128, email: takis@ee.tamu.edu.

0-7803-9098-9/05/$25.00 ©2005 AACC

system and to drive $e_o(t)$ defined in (1) close to zero. We

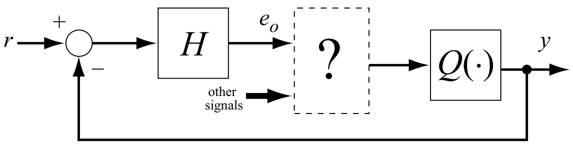

Fig. 3. Casting of the oversampled encoding problem as a tracking-control problem.

note that if the size of e_o is made sufficiently small, the (discretized) output signal y represents the signal r to within any desired accuracy. In the following section, we address issues of closed-loop stability/tracking and the design of the controller.

III. Novel Approach

In the following, vectors and matrices are indicated by bold characters. The notation $\|\cdot\|$ denotes the Euclidean norm of a vector or the induced norm of a matrix. The set B_x represents the open ball of radius $x \in \mathbb{R}$, $x > 0$, centered at the origin, i.e., $B_x := \{\boldsymbol{e} \in \mathbb{R}^n : \|\boldsymbol{e}\| < x\}$. Our development that follows is in continuous-time.

The proposed architecture is shown in block diagram form in Figure 4, with controller design shown in Figure 5. The architecture was invented through the application of variable-structure methods [7] to the tracking problem posed in Figure 3.

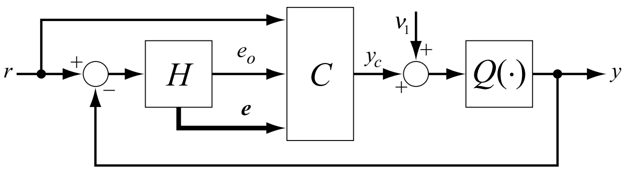

Fig. 4. Proposed converter architecture.

The input r and disturbance inputs ν_1 and ν_2 are real, scalar quantities. The filter H, introduced in (1), is described by the state model:

$$H : \begin{cases} \dot{\boldsymbol{e}} = \boldsymbol{A}\boldsymbol{e} + \boldsymbol{b}(r - y) \\ e_o = \boldsymbol{c}\boldsymbol{e} \end{cases} \tag{2}$$

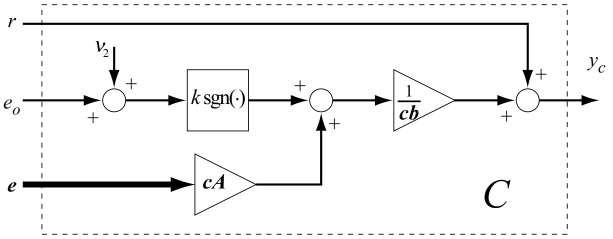

Fig. 5. Proposed nonlinear controller, C, to regulate the error, e_o. The comparator element, $k\,\mathrm{sgn}(\cdot)$, represents a switching-type nonlinearity.

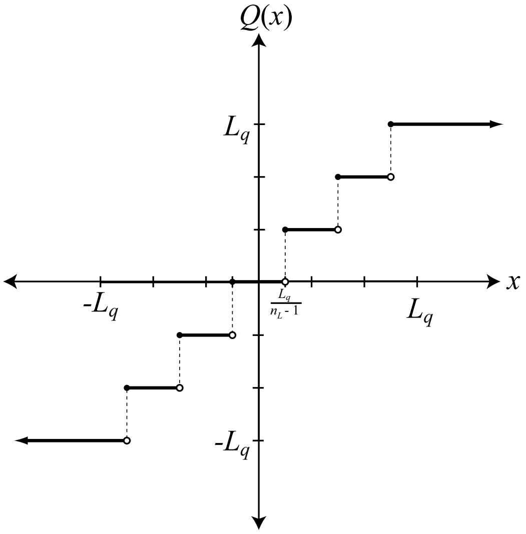

Fig. 6. Illustration of quantizer characteristic assumed in this paper (7-level quantizer shown).

where $\boldsymbol{e} \in \mathbb{R}^n$, $n \in \mathbb{Z}$ $(n > 0)$, e_o, $y \in \mathbb{R}$. The matrices $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are of appropriate dimensions with constant real entries. An initial state $\boldsymbol{e}(0) \in \mathbb{R}^n$ accompanies the state equations.

The controller C is described by

$$C : \; y_c = r + (\boldsymbol{cb})^{-1} \left[\boldsymbol{cAe} + k\,\mathrm{sgn}(e_o + \nu_2)\right] \tag{3}$$

in which the comparator gain k is provided in (6), below. The output of the architecture is given by

$$y = Q(y_c + \nu_1) \tag{4}$$

where $Q(\cdot)$ is a quantizer with n_L levels and outputs limits $\pm L_q$, $0 < L_q < \infty$, as shown in Figure 6. We may also write

$$y = y_c + \nu_1 + w_Q(y_c + \nu_1) \tag{5}$$

where $w_Q(x) := Q(x) - x$, $x \in \mathbb{R}$.

We provide two theorems. The first theorem applies to the foregoing system description assuming bounded disturbances ν_1 and ν_2. In the second theorem, ν_1 and ν_2 are used to model the sampling operations that occur in practice immediately preceding the quantizer, $Q(\cdot)$, and comparator, $k\,\mathrm{sgn}(\cdot)$.[1]

We make the standing assumptions:

- *Assumption 1 (bounded input):* The input signal $r(t)$ is bounded, i.e., $\sup_{t \geq 0} |r(t)| \leq \infty$.
- *Assumption 2 (constraints on H):* The matrix $\boldsymbol{A}$ is Hurwitz, i.e., all eigenvalues of $\boldsymbol{A}$ have negative real parts, and each zero of H has a negative real part. Moreover, we assume that the state-space model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, 0)$ is controllable and that $\boldsymbol{cb} \neq 0$.

[1] In modern analog VLSI systems, comparators are often “track-and-latch” designs in which a sampling operation is implicit. We shall assume that the triggering of the quantizer and comparator is synchronized by a common clock which determines the sampling rate of the system.

We have the following definitions:

- *Definition 1:* The architecture of Figures 4 and 5 is said to be **stable** if:
  1) $\|\boldsymbol{e}(t)\| < \infty$ for all $t \geq 0$, and
  2) there exists a $T_\delta > 0$ such that $\boldsymbol{e}(t) \in B_\delta$ for all $t \geq T_\delta$, for some δ, $0 < \delta < \infty$,

  for any initial condition $\boldsymbol{e}(0)$ within a bounded domain $D \subset \mathbb{R}^n$ and for any r, ν_1 and ν_2 within specified bounds.
- *Definition 2:* $M_r \in \mathbb{R}$ is a known constant such that $\sup_{t \geq 0} |r(t)| \leq M_r$.
- *Definition 3:* $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ is the symmetric, positive definite solution to $\boldsymbol{PA} + \boldsymbol{A}^T\boldsymbol{P} = -\boldsymbol{I}$.
- *Definition 4:* $\mu := 2(M_r + L_q)\|\boldsymbol{b}\|\lambda_{\text{max}}(\boldsymbol{P}) + \varepsilon_\mu$ for some $\varepsilon_\mu > 0$, where $\lambda_{\text{max}}(\boldsymbol{P})$ denotes the largest eigenvalue of $\boldsymbol{P}$.
- *Definition 5:* $\rho := \left(\sqrt{\frac{\lambda_{\text{max}}(\boldsymbol{P})}{\lambda_{\text{min}}(\boldsymbol{P})}}\right)\mu + \varepsilon_\rho$ for some $\varepsilon_\rho > 0$ where $\lambda_{\text{min}}(\boldsymbol{P})$ denotes the smallest eigenvalue of $\boldsymbol{P}$.
- *Definition 6:* $\varepsilon := \lambda_{\text{max}}(\boldsymbol{P})\mu^2$, $\xi := \lambda_{\text{min}}(\boldsymbol{P})\rho^2$.
- *Definition 7:* $\bar{\boldsymbol{A}} := (\boldsymbol{I} - \boldsymbol{b}(\boldsymbol{cb})^{-1}\boldsymbol{c})\boldsymbol{A}$.
- *Definition 8:* $\boldsymbol{\Psi}(t) := \exp(\bar{\boldsymbol{A}}t)$.
- *Definition 9:*
  $N_\psi := \sup_{t \geq 0} \|\boldsymbol{\Psi}(t)\| + \|\boldsymbol{b}(\boldsymbol{cb})^{-1}\|(\|\boldsymbol{\Psi}(0)\| + \sup_{t \geq 0} \|\boldsymbol{\Psi}(t)\| + \sup_{t \geq 0} \int_0^t \|\dot{\boldsymbol{\Psi}}(\tau)\| \, d\tau)$.
- *Definition 10:* $\Omega_\varepsilon := \{\boldsymbol{e} \in \mathbb{R}^n : \boldsymbol{e}^T\boldsymbol{Pe} \leq \varepsilon\}$.
- *Definition 11:* $\Omega_\xi := \{\boldsymbol{e} \in \mathbb{R}^n : \boldsymbol{e}^T\boldsymbol{Pe} \leq \xi\}$.

Definition 1 concerns stability and tracking and is akin to boundedness with ultimate boundedness as described by Khalil [8]. The size of δ is addressed in the main result of this development; it is a function of the magnitude of ν_2 and the characteristics of the filter H. Also note that since $\boldsymbol{A}$ is Hurwitz, a positive definite $\boldsymbol{P}$ can always be found.

Theorem 1: We assume that $\sup_{t \geq 0} |\nu_1(t)| < \infty$ and $\sup_{t \geq 0} |\nu_2(t)| < \infty$. Given M_{ν_1}, $M_{\nu_2} \in \mathbb{R}$ such that $\sup_{t \geq 0} |\nu_1(t)| \leq M_{\nu_1}$, $\sup_{t \geq 0} |\nu_2(t)| \leq M_{\nu_2}$ and the following additional assumptions:

- *Assumption 3:* $N_\psi M_{\nu_2} < \mu$.
- *Assumption 4:*
  $2M_{\nu_1} + M_r + |(\boldsymbol{cb})^{-1}|\|\boldsymbol{cA}\|\sqrt{\frac{\varepsilon}{\lambda_{\text{min}}(\boldsymbol{P})}} = L_q - \varepsilon_q$, where ε_q is a known positive finite constant.

Then the architecture shown in Figures 4 and 5 described as above with controller gain k given by:

$$k = |\boldsymbol{cb}|\left(M_{\nu_1} + \frac{L_q}{n_L - 1}\right) + \varepsilon_k \tag{6}$$

where $\varepsilon_k \geq \varepsilon_q$ is stable in the sense of Definition 1 where the domain $D = \Omega_\xi$. Furthermore, $\delta = \delta(t) \to N_\psi M_{\nu_2}$ as $t \to \infty$.◇

The proof of Theorem 1 has two parts. Part 1) establishes that $\boldsymbol{e}(t)$ must remain within Ω_ξ and enters the set $\Omega_\varepsilon \subset \Omega_\xi$ within finite time (from any initial condition $\boldsymbol{e}(0)$ within Ω_ξ). Part 2) addresses the sliding mode behavior of the system, showing that, after entering Ω_ε, $\boldsymbol{e}(t)$ is ultimately restricted to a neighborhood of the origin with radius given by $N_\psi M_{\nu_2}$.

Proof of Theorem 1: (sketch)

1) Here we show that for any $\boldsymbol{e}(0) \in \Omega_\xi$, there exists a time $T_\varepsilon > 0$ such that $\boldsymbol{e}(t) \in \Omega_\varepsilon$ for all $t \geq T_\varepsilon$. Consider the positive definite function

$$V_a = \boldsymbol{e}^T\boldsymbol{Pe} \tag{7}$$

and suppose that $\boldsymbol{e}(0) \in \Omega_\xi$. Note that $\lambda_{\text{min}}(\boldsymbol{P})\|\boldsymbol{e}\|^2 \leq V_a \leq \lambda_{\text{max}}(\boldsymbol{P})\|\boldsymbol{e}\|^2$. Taking time derivatives,

$$\begin{aligned}\dot{V}_a = \dot{\boldsymbol{e}}^T\boldsymbol{Pe} + \boldsymbol{e}^T\boldsymbol{P}\dot{\boldsymbol{e}} = \boldsymbol{e}^T(\boldsymbol{PA} + \boldsymbol{A}^T\boldsymbol{P})\boldsymbol{e} +& \\ 2(r - y)\boldsymbol{b}^T\boldsymbol{Pe}.&\end{aligned} \tag{8}$$

Since $\boldsymbol{PA} + \boldsymbol{A}^T\boldsymbol{P} = -\boldsymbol{I}$, we see that

$$\begin{aligned}\dot{V}_a &= -\|\boldsymbol{e}\|^2 + 2(r - y)\boldsymbol{b}^T\boldsymbol{Pe} \\ &\leq -\|\boldsymbol{e}\|[\|\boldsymbol{e}\| - 2(M_r + L_q)\|\boldsymbol{b}\|\lambda_{\text{max}}(\boldsymbol{P})].\end{aligned} \tag{9}$$

Therefore, $\dot{V}_a < 0$ if $\|\boldsymbol{e}\| \geq \mu$. So we have that $\dot{V}_a < 0$ for all $\boldsymbol{e}$ outside the open ball B_μ. The set Ω_ε is a level set of V_a containing B_μ since

$$\|\boldsymbol{e}\| \leq \mu \Rightarrow \lambda_{\text{max}}(\boldsymbol{P})\|\boldsymbol{e}\|^2 \leq \lambda_{\text{max}}(\boldsymbol{P})\mu^2 = \varepsilon, \tag{10}$$

which implies that $V_a = \boldsymbol{e}^T\boldsymbol{Pe} \leq \lambda_{\text{max}}(\boldsymbol{P})\|\boldsymbol{e}\|^2 \leq \varepsilon$. In addition, $\dot{V}_a < 0$ on the boundary of Ω_ε (since $V_a = \varepsilon \Rightarrow \lambda_{\text{max}}(\boldsymbol{P})\|\boldsymbol{e}\|^2 \geq \varepsilon \Rightarrow \|\boldsymbol{e}\| \geq \mu \Rightarrow \dot{V}_a < 0$). Since V_a is strictly decreasing for $\boldsymbol{e}$ outside and on the boundary of Ω_ε, and Ω_ξ is a level set containing Ω_ε, we see that $\boldsymbol{e}(t)$ must enter Ω_ε from any initial point $\boldsymbol{e}(0)$ in Ω_ξ in finite time T_ε, $0 \leq T_\varepsilon < \infty$. Moreover, $\boldsymbol{e}(t)$ can escape neither Ω_ε nor Ω_ξ since V_a is strictly decreasing for $\|\boldsymbol{e}\| \geq \mu$. Hence we have also established that $\|\boldsymbol{e}(t)\| < \rho$ for all $t \geq 0$.

2) Here we demonstrate that as $t \to \infty$, $\delta(t) \to N_\psi M_{\nu_2}$ (where δ is described in Definition 1). We define

$$\Lambda_{M_{\nu_2}} := \{\boldsymbol{e} \in \mathbb{R}^n : |e_o| = |\boldsymbol{ce}| \leq M_{\nu_2}\}. \tag{11}$$

We note that

$$|w_Q(\nu_1 + y_c)| \leq \left\{ \begin{array}{l} \frac{L_q}{n_L - 1}, 0 \leq |\nu_1 + y_c| < \bar{L}_q \\ |\nu_1 + y_c| - L_q, |\nu_1 + y_c| \geq \bar{L}_q \end{array} \right.$$

where $\bar{L}_q := L_q\left(1 + \frac{1}{n_L - 1}\right)$. Also note that for $\boldsymbol{e} \in \Omega_\varepsilon$ (which is the case for all $t \geq T_\varepsilon$), $\|\boldsymbol{e}\| \leq \sqrt{\frac{\varepsilon}{\lambda_{\text{min}}(\boldsymbol{P})}}$, since $\boldsymbol{e}^T\boldsymbol{Pe} \geq \lambda_{\text{min}}(\boldsymbol{P})\|\boldsymbol{e}\|^2$. Hence,

$$\begin{aligned}|\nu_1 + y_c| - L_q \leq M_{\nu_1} + M_r +& \\ |(\boldsymbol{cb})^{-1}|k + |(\boldsymbol{cb})^{-1}|\|\boldsymbol{cA}\|\sqrt{\frac{\varepsilon}{\lambda_{\text{min}}(\boldsymbol{P})}} - L_q.&\end{aligned} \tag{12}$$

Now consider the scalar function

$$V = \frac{1}{2}e_o^2. \tag{13}$$

Differentiating with respect to time, we obtain $\dot{V} = e_o\dot{e}_o = e_o\boldsymbol{cAe} + e_o(\boldsymbol{cb})r - e_o(\boldsymbol{cb})y$. We have that $y = \nu_1 + \left[r + (\boldsymbol{cb})^{-1}k\,\text{sgn}(e_o + \nu_1) + (\boldsymbol{cb})^{-1}\boldsymbol{cAe}\right] + w_Q(\nu_1 + y_c)$. Therefore, $\dot{V} = -k\,\text{sgn}(e_o + \nu_2)e_o -$

$(\boldsymbol{cb})\nu_1 e_o - (\boldsymbol{cb}) w_Q(\nu_1 + y_c) e_o$. Suppose $|e_o| \geq M_{\nu_2}$. Thus we have

$$\dot{V} \leq -k|e_o| + |\boldsymbol{cb}| M_{\nu_1} |e_o| + \\ |\boldsymbol{cb}||w_Q(\nu_1 + y_c)||e_o|. \tag{14}$$

We now consider two cases. In the first instance, $0 \leq |\nu_1 + y_c| < L_q \left(1 + \frac{1}{n_L - 1}\right)$, so that for $|e_o| \geq M_{\nu_2}$. In the second case, $|\nu_1 + y_c| \geq L_q \left(1 + \frac{1}{n_L - 1}\right)$. Using Assumption 4, and setting $k = |\boldsymbol{cb}| \left(M_{\nu_1} + \frac{L_q}{n_L - 1}\right) + \varepsilon_k$ ensures that $\dot{V} \leq -\varepsilon_q |e_o|$, for $|\nu_1 + y_c| \geq 0$, $\boldsymbol{e} \in \Omega_\varepsilon$, $|e_o| \geq M_{\nu_2}$. Thus there exists a finite time $T_{M_{\nu_2}} > T_\varepsilon$ such that $\boldsymbol{e}(t) \in \Omega_{M_{\nu_2}}$ for all $t \geq T_{M_{\nu_2}}$. Once inside $\Lambda_{M_{\nu_2}}$, $\boldsymbol{e}(t)$ cannot escape from $\Lambda_{M_{\nu_2}} \bigcap \Omega_\varepsilon$.

The following is adapted from Utkin [7] and concerns the evolution of $\boldsymbol{e}(t)$ within $\Lambda_{M_{\nu_2}}$. Using the method of equivalent control and integration-by-parts, we obtain the solution of the error dynamics restricted to the set $\Lambda_{M_{\nu_2}}$ as

$$\begin{aligned} \boldsymbol{e}(t) = \, & \boldsymbol{\Psi}(t - T_{M_{\nu_2}}) \boldsymbol{e}(T_{M_{\nu_2}}) + \\ & \boldsymbol{\Psi}(0) \boldsymbol{b} (\boldsymbol{cb})^{-1} e_o(t) - \\ & \boldsymbol{\Psi}(t - T_{M_{\nu_2}}) \boldsymbol{b} (\boldsymbol{cb})^{-1} e_o(T_{M_{\nu_2}}) - \\ & \int_{T_{M_{\nu_2}}}^{t} \dot{\boldsymbol{\Psi}}(t - \tau) \boldsymbol{b} (\boldsymbol{cb})^{-1} e_o(\tau) \, d\tau, \end{aligned} \tag{15}$$

where $\boldsymbol{\Psi} : \mathbb{R} \mapsto \mathbb{R}^{n \times n}$ denotes the state transition matrix. We now define the sliding mode system, $\boldsymbol{\Sigma}_{\text{sm}}$ as the dynamics $\boldsymbol{\Sigma}_{e,\, \Lambda_{M_{\nu_2}}}$ restricted to the switching surface $\boldsymbol{ce} = 0$ as

$$\boldsymbol{\Sigma}_{\text{sm}} : \begin{cases} \dot{\bar{\boldsymbol{e}}} = \bar{\boldsymbol{A}} \bar{\boldsymbol{e}} \\ \boldsymbol{c} \bar{\boldsymbol{e}} = 0 \end{cases}, \; t \geq T_{M_{\nu_2}} \tag{16}$$

with initial condition

$$\bar{\boldsymbol{e}}(T_{M_{\nu_2}}) = \left[\boldsymbol{e}(T_{M_{\nu_2}}) - M_{\nu_2} \frac{\boldsymbol{c}^T}{\|\boldsymbol{c}\|} \right]. \tag{17}$$

Thus the solution to (16) can be written as

$$\bar{\boldsymbol{e}}(t) = \boldsymbol{\Psi}(t - T_{M_{\nu_2}}) \bar{\boldsymbol{e}}(T_{M_{\nu_2}}). \tag{18}$$

It can be shown from our assumption of controllability and the fact that all zeros of H are in the open left-half plane that (16) is exponentially stable (please see [5] for details). Subtracting (18) from (15), taking norms of both sides and noting that $\sup_{\forall t \geq T_{M_{\nu_2}}} |e_o(t)| = M_{\nu_2}$ we have

$$\begin{aligned} \|\boldsymbol{e}(t) - \bar{\boldsymbol{e}}(t)\| \leq \, & \|\boldsymbol{\Psi}(t - T_{M_{\nu_2}})\| M_{\nu_2} + \\ & \|\boldsymbol{b}(\boldsymbol{cb})^{-1}\| \Big[\|\boldsymbol{\Psi}(0)\| + \\ & \|\boldsymbol{\Psi}(t - T_{M_{\nu_2}})\| + \\ & \int_0^{t - T_{M_{\nu_2}}} \|\dot{\boldsymbol{\Psi}}(t - \tau)\| d\tau \Big] M_{\nu_2}. \end{aligned} \tag{19}$$

Therefore,

$$\|\boldsymbol{e}(t) - \bar{\boldsymbol{e}}(t)\| \leq N_\psi M_{\nu_2}, \; t \geq T_{M_{\nu_2}}. \tag{20}$$

Since $\boldsymbol{\Sigma}_{\text{sm}}$ is stable, the state transition matrix $\boldsymbol{\Psi}(t)$ and the integral $\int_0^t \|\dot{\boldsymbol{\Psi}}(\tau)\| \; d\tau$ are bounded. Thus, $N_\psi < \infty$. Using (20), we obtain $\|\boldsymbol{e}(t)\| \leq N_\psi M_{\nu_2} + \|\bar{\boldsymbol{e}}(t)\|$, $t \geq M_{\nu_2}$, which shows that $\boldsymbol{e}(t)$ is restricted to the open ball $B_{\delta(t)}$ where

$$\begin{aligned} \lim_{t \to \infty} \delta(t) & = \lim_{t \to \infty} [N_\psi M_{\nu_2} + \|\bar{\boldsymbol{e}}(t)\|] \\ & = N_\psi M_{\nu_2}. \end{aligned} \tag{21}$$

□

We now extend Theorem 1 to account for sample-and-hold operations occurring at the quantizer and comparator elements, in the manner shown in Figure 7. Note that Figure 8 provides an equivalent representation for Figure 7, where $e_x(t) = x_1(t) - x_1(n_t T_s)$ and $n_t := \left\lfloor \frac{t}{T_s} \right\rfloor$, where $\lfloor \cdot \rfloor$ denotes the greatest integer less than its argument.

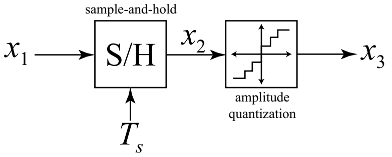

Fig. 7. Sampling operation associated with quantizer and comparator; x_2 is given by $x_2(t) = x_1(n_t T_s)$, in which $n_t := \left\lfloor \frac{t}{T_s} \right\rfloor$.

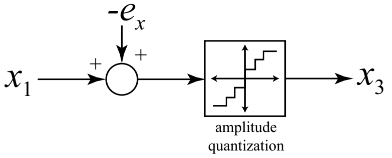

Fig. 8. Alternative representation for sampling, where $e_x(t) = x_1(t) - x_1(n_t T_s)$.

Theorem 2: Given that Assumptions 1 and 2 hold, that M_{ν_1} and M_{ν_2} are any numbers in the interval $(0, \infty)$ consistent with Assumptions 3 and 4 from the statement of Theorem 1, and that the following hold:

- *Assumption 5:* the sample-and-hold operations associated with the quantizer and the comparator are triggered by a common clock with a fixed period, $T_s > 0$.
- *Assumption 6:*

$$\begin{aligned} \nu_1(t) & := -(y_c(t) - y_c(n_t T_s)), \\ \nu_2(t) & := -(e_o(t) - e_o(n_t T_s)). \end{aligned}$$

- *Assumption 7:* there exists a known $M_{\dot{r}} \in (0, \infty)$, such that $\sup_{t \geq 0} |\dot{r}(t)| \leq M_{\dot{r}}$.
- *Assumption 8:* the sampling period T_s satisfies

$$\begin{aligned} 0 < T_s \leq \min \Big(& T_s', \; M_{\nu_2} (\|\boldsymbol{c}\| M_{T_s})^{-1}, \\ & M_{\nu_1} \left[M_{\dot{r}} + \|(\boldsymbol{cb})^{-1} \boldsymbol{cA}\| M_{T_s} \right]^{-1} \Big), \end{aligned}$$

  where T_s' and M_{T_s} are defined in the proof of Theorem 2.

Then the architecture of Figures 4 and 5 described as above with controller gain k as defined in the statement of Theorem 1 is stable in the sense of Definition 1.◇

This theorem establishes that, for sufficiently high sample rates, stability is guaranteed. Note that a finite rate of change for r can be guaranteed through filtering, therefore the restriction on the rate of change of r is not impractical.

Proof of Theorem 2 (sketch):

Since $\boldsymbol{A}$ is Hurwitz, H is bounded-input, bounded-output stable, i.e., it can be shown that $\|\boldsymbol{e}(t)\| \leq \alpha\|\boldsymbol{e}(0)\| + \beta\|\boldsymbol{b}\|L_q =: M_{\boldsymbol{e}} < \infty$, for some positive finite real constants α and β. Therefore, we can show that for any $\varepsilon_s > 0$ and for any $t \geq 0$, there exists a $T_s' > 0$ (independent of t) such that

$$\|\boldsymbol{e}(t) - \boldsymbol{e}(n_tT_s)\| \leq \Big[(\|\boldsymbol{A}\| + \varepsilon_s)M_{\boldsymbol{e}} + 2L_q\|\boldsymbol{b}\| \sup_{t\geq 0} \|\exp(-\boldsymbol{A}t)\|\Big] T_s, \tag{22}$$

for all $t \geq 0$ and for all $0 < T_s \leq T_s'$, where

$$M_{T_s} := \Big[(\|\boldsymbol{A}\| + \varepsilon_s)M_{\boldsymbol{e}} + 2L_q\|\boldsymbol{b}\| \sup_{t\geq 0} \|\exp(-\boldsymbol{A}t)\|\Big] < \infty. \tag{23}$$

Consider the sampling operation at the comparator, where we obtain

$$\begin{aligned} |\nu_2(t)| &= |e_o(t) - e_o(n_tT_s)| \\ &= |\boldsymbol{ce}(t) - \boldsymbol{ce}(n_tT_s)| \leq \|\boldsymbol{c}\|M_{T_s}T_s, \end{aligned} \tag{24}$$

for all $t \geq 0$. Thus, to maintain $|e_o(t) - e_o(n_tT_s)| \leq M_{\nu_2}$ for all $t \geq 0$, it is sufficient to have $0 < T_s \leq M_{\nu_2}(\|\boldsymbol{c}\|M_{T_s})^{-1}$. For the sampling operation at the quantizer, we have

$$\begin{aligned} |\nu_1(t)| &= |y_c(t) - y_c(n_tT_s)| \\ &= \big|r(t) - r(n_tT_s) + (\boldsymbol{cb})^{-1}\boldsymbol{cA}[\boldsymbol{e}(t) - \boldsymbol{e}(n_tT_s)]\big| \end{aligned} \tag{25}$$

where we have used the fact that within any given sampling interval $[n_tT_s, (n_t+1)T_s)$, $\operatorname{sgn}[e_o(t) + \nu_2(t)] = \operatorname{sgn}[e_o(n_tT_s) + \nu_2(n_tT_s)]$. Thus, using the fact that the time-derivative of $r(t)$ is bounded, we find

$$|y_c(t) - y_c(n_tT_s)| \leq \Big[M_{\dot{r}} + \|(\boldsymbol{cb})^{-1}\boldsymbol{cA}\|M_{T_s}\Big]T_s. \tag{26}$$

Therefore, to ensure that $|y_c(t) - y_c(n_tT_s)| \leq M_{\nu_1}$ for all $t \geq 0$, it is sufficient to have $0 < T_s \leq M_{\nu_1}\Big[M_{\dot{r}} + \|(\boldsymbol{cb})^{-1}\boldsymbol{cA}\|M_{T_s}\Big]^{-1}$.□

IV. Simulations

We empirically study the quantization-noise-shaping characteristics of the proposed encoding scheme. We examine the spectral linearity of the architecture within the signal band and assess dynamic range. Our principal metric of performance is the signal-to-quantization-noise ratio (SQNR), measured in the frequency domain. The SQNR is defined as the ratio beween signal power to quantization error power within the signal band.

We employ an encoder design with multi-level quantization that operates at a moderate oversampling ratio, consistent with modern applications requiring approximately 1 MHz of signal bandwidth and in excess of 14 bits of dynamic range [9]. A 5th-order 3-bit (9-level) version of the architecture is used operating at an OSR (denoted by ζ) of $\zeta = 32$ and sampling rate (denoted by f_s) of $f_s = 64$ MHz. The upper limit of the signal band is denoted by f_b so that the signal band extends from dc to $f_b := \frac{f_s}{2\zeta}$. An ideal quantizer is modeled. The transfer function of H is $H(s) = \frac{p(s)}{q(s)}$, where $p(s) = 9264s^4 + (1.779 \times 10^{11})s^3 + (7.259 \times 10^{18})s^2 + (6.149 \times 10^{25})s + 9.384 \times 10^{32}$ and $q(s) = s^5 + (4.424 \times 10^6)s^4 + (6.074 \times 10^{13})s^3 + (1.79 \times 10^{20})s^2 + (7.807 \times 10^{26})s + 9.384 \times 10^{32}$. The quantizer output limit parameter, $L_q = 1$. The simulations for this section are based on a fixed step size of $\frac{T_s}{32}$, where $T_s = \frac{1}{f_s}$; each simulation consists of 32768 samples. The controller gain parameter used is $k = 1.05|\boldsymbol{cb}|$.

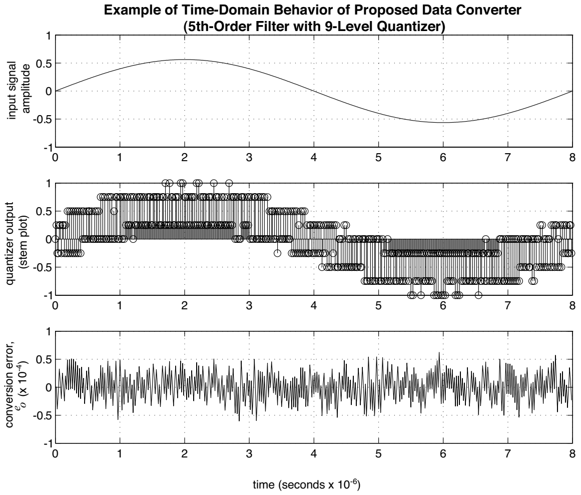

Fig. 9. Sample of time-domain behavior of converter input signal (top plot), quantizer output (middle plot) and conversion error e_o (bottom plot); one period of the input signal is shown. We confirm that the converter output (middle plot) is a digital signal, discrete in both amplitude and time. The conversion error is the output of the filter H.

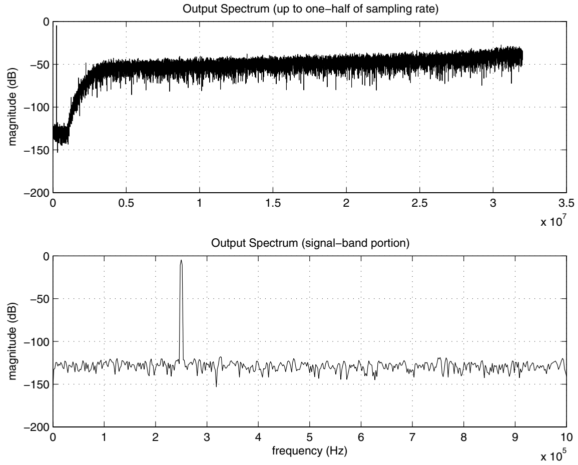

Fig. 10. Single-tone test; input tone at a frequency of $\frac{f_b}{4}$. Output spectrum up to $\frac{f_s}{2}$ (top plot); signal-band portion of output spectrum (bottom plot). The SQNR is 96.7 dB.

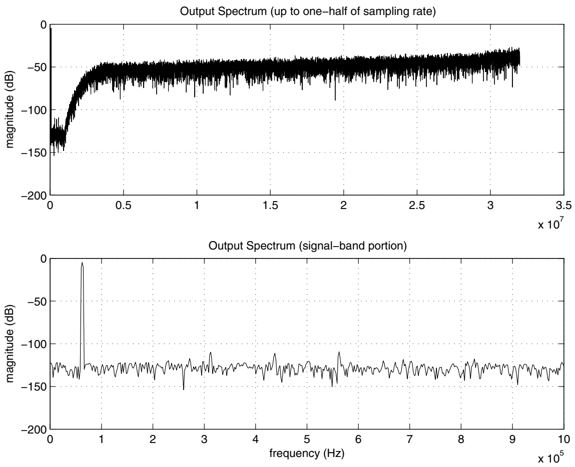

Fig. 11. Single-tone test; input tone at a frequency of $\frac{f_b}{16}$. Output spectrum up to $\frac{f_s}{2}$ (top plot); signal-band portion of output spectrum (bottom plot). The SQNR is 95.2 dB.

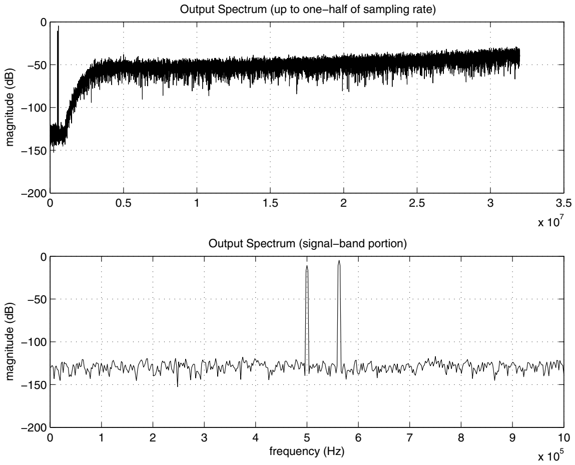

Fig. 12. Two-tone test; equal amplitude input tones at frequencies of $\frac{f_b}{2}$ and $\frac{9f_b}{16}$. Output spectrum up to $\frac{f_s}{2}$ (top plot); signal-band portion of output spectrum (bottom plot). The SQNR is 95.7 dB.

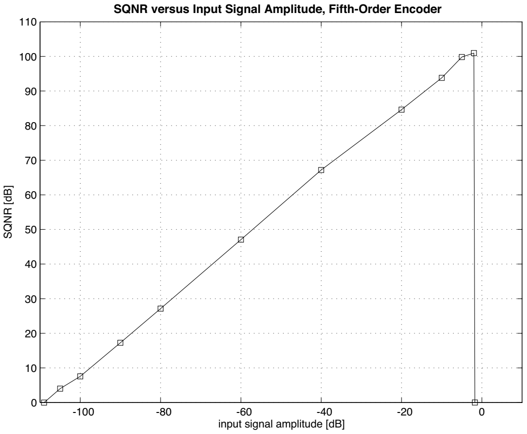

Fig. 13. Plot of SQNR versus input signal amplitude from simulation data. Note that an input signal amplitude of 0 dB corresponds to an amplitude of $L_q = 1$.

To begin, we provide a representative time-domain sample of the converter behavior in Figure 9.

A single-tone test yields the output shown in Figure 10, where harmonic distortion is not apparent within the signal band. The noise floor is essentially flat from dc up to $f = f_b$. Note that, in contrast, delta-sigma modulated spectra often exhibit notches in the noise floor. Varying the frequency of this tone within the signal band produces similar, highly linear responses, as seen in Figure 11. A two-tone test provides the output spectrum in Figure 12. The sum and difference frequencies, markers of intermodulation distortion, are not apparent above the noise floor.

The SQNR versus input signal power curve for the 5th-order encoder is shown in Figure 13. This test gives a sense of the performance across a broad range of input levels and provides a means of evaluating dynamic range. From Figure 13, we see that the 5th-order converter exhibits a dynamic range of approximately 110 dB. We see a linear (slope of 1) progression of SQNR with input signal amplitude (measured in decibels), which is a desirable characteristic in data converters.

V. Conclusion

We propose a novel oversampled conversion technique that provides high levels of SQNR performance and that is based on a nonlinear control strategy for the explicit reduction of a performance error (denoted by e_o). Our principal contributions are: 1) the formulation of the data conversion problem as a tracking-control problem, 2) the interpretation of the states of the performance filter H as a set of *error dynamics*, and 3) the use of variable-structure techniques for the design of a controller to stabilize the error dynamics.

References

[1] I. Galton, "Delta-sigma data conversion in wireless transceivers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, pp. 302–315, 2002.

[2] P. M. Aziz, H. V. Sorensen, and J. van der Spiegel, "An overview of sigma-delta converters," *IEEE Signal Processing Magazine*, vol. 13, pp. 61–84, 1996.

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

[4] 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. Norsworthy, R. Schreier, and G. Temes, Eds. IEEE Press, 1997, pp. 447–467.

[5] T. Zourntos and D. A. Johns, "Variable-structure compensation of delta-sigma modulators: Stability and performance," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, pp. 41–53, 2002.

[6] S. Plekhanov, I. A. Shkolnikov, and Y. B. Shtessel, "High order sigma-delta modulator design via sliding mode control," in *Proceedings of the American Control Conference*, 2003, pp. 897–902.

[7] V. I. Utkin, *Sliding Modes in Control and Optimization*. Springer-Verlag, 1992.

[8] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice-Hall, 2002.

[9] K. Vleugels, S. Rabii, and B. A. Wooley, "A 2.5-V sigma-delta modulator for broadband communications applications," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1887–1899, 2001.