Spectral Characteristics of the Double-Loop Sigma–Delta Modulator

Mariam Motamed, Avideh Zakhor, Seth Sanders, and Te-Won Lee

***Abstract*— An important factor affecting the performance of $\Sigma\Delta$ modulators is their tone behavior. A recent approach to alleviate the tone problem consists of moving the open-loop poles outside the unit circle. In this paper, we determine the frequency location of two dominant tones as a function of pole-location, and discuss the effect of pole location on the relative amplitude of low-frequency tones. Audio testing along with a series of simulations are then performed to compare the efficacy of pole placement with that of the more traditional tone removal technique of dithering.**

***Index Terms*—Double loop sigma delta modulator, tone behavior.**

I. Introduction

Sigma–delta $(\Sigma\Delta)$ modulators are becoming increasingly important in data conversion applications. An important factor limiting the performance of $\Sigma\Delta$ modulators is their tone behavior [1], [2]. The traditional tone removal technique consists of using additive dither signals to randomize the bit pattern at the output of $\Sigma\Delta$ modulators [2]. A more recent approach to alleviate the tone problem is to design $\Sigma\Delta$ modulators with open-loop poles outside the unit circle [3]–[5].

In this paper, we consider the tonal properties of the double-loop $\Sigma\Delta$ modulator with unstable filter dynamics,[1] and compare them to those of the dithered double-loop $\Sigma\Delta$ modulator. Audio testing has also been performed in order to evaluate audible distortion as a function of tone amplitude. The paper is organized as follows. In Section II, we obtain an estimate of the average value of the output bit sequence as a function of the system parameters. In Section III, we determine the frequency location of two dominant spectral components as a function of pole location, and show via simulations that pole values slightly outside the unit circle are more advantageous in suppressing low-frequency tones than large poles outside the unit circle. In Section IV, we present the results of our audio testing, along with a comparison between the dithered modulator and the modulator with open-loop poles outside the unit circle. In all that follows, our analysis and simulations assume constant encoder input.

II. Output Sample Mean

Consider the double-loop $\Sigma\Delta$ modulator shown in Fig. 1, where p_1 and p_2 are the integrator poles, $Q(\cdot)$ is a 1-bit quantizer, signals $u(\cdot)$ and $v(\cdot)$ denote the states of the integrators, X is the dc input, and $y(\cdot)$ is the output sequence. Throughout this section and Section III, we assume that signal $d(\cdot)$ in Fig. 1 is identically zero. Let $\epsilon(\cdot)$ denote the quantization error sequence $\epsilon(n) = y(n) - u(n)$. The Z transform of $y(\cdot)$, $Y(z)$ can be written in terms of the Z transforms of X, $X(z)$, and $\epsilon(\cdot)$, $\varepsilon(z)$ as follows:

$$Y(z) = G_1(z)X(z) + G_2(z)\varepsilon(z) \tag{1}$$

where G_1 corresponds to the signal transfer function $G_1(z) = z^{-1}/[1+(-p_1-p_2+2)z^{-1}+(p_1p_2-p_1)z^{-2}]$, and G_2 corresponds to the quantization noise transfer function $G_2(z) = [(1-p_1z^{-1})(1-p_2z^{-1})]/[1+(-p_1-p_2+2)z^{-1}+(p_1p_2-p_1)z^{-2}]$. Evaluating (1) at dc yields $Y(1) = G_1(1)X(1) + G_2(1)\varepsilon(1)$. Using Taylor series expansion and neglecting second and higher order terms, we obtain simplified expressions for $G_1(1)$ and $G_2(1)$ for the case when both poles are near 1:

$$G_1(1) \approx p_1 \quad \text{and} \quad G_2(1) \approx 0. \tag{2}$$

Thus, for real pole values close to 1, $Y(1) \approx p_1X(1)$. This approximation is used in Section III-B to evaluate the sample mean, $\hat{m}_y$, of the output bit sequence, $\hat{m}_y \triangleq (1/N)\ \Sigma_{n=0}^{N-1}\ y(n) = (1/N)Y(1)$.

III. Spectral Characteristics

In this section, we consider the spectral characteristics of the double-loop modulator for two cases. We begin in Section III-A by considering the case where p_2 is fixed on the unit circle while p_1 is allowed to move outside the unit circle.[2] We consider the case where both poles are outside the unit circle in Section III-B.

A. Case 1: $p_1 > 1$ and $p_2 = 1$

Assume that the output bit sequence $y(\cdot)$ is a periodic sequence with period N. Let N_1 and N_2 denote the number of $+1$'s and -1's, respectively, in N, such that $N = N_1 + N_2$ and $\hat{m}_y \triangleq (1/N)(N_1 - N_2)$. Solving for N_1 and N_2 as a function of N and $\hat{m}_y$ yields

$$\begin{aligned} N_1 &= \tfrac{1}{2}N(1+\hat{m}_y) \quad N_1 \in \mathcal{N}^+ \\ N_2 &= \tfrac{1}{2}N(1-\hat{m}_y) \quad N_2 \in \mathcal{N}^+ \end{aligned} \tag{3}$$

where $\mathcal{N}^+$ denotes the set of positive integers, and N is any integer that yields integer values for both N_1 and N_2. Since $p_2 = 1$ and $G_2(1) = 0$, the expression for $\hat{m}_y$ used to evaluate N_1 and N_2 in (3) is $\hat{m}_y = G_1(1)X$.

If the output sequence $y(\cdot)$ is periodic with period N, then N must satisfy the equalities in (3) in the sense that the resulting N_1 and N_2 as computed in (3) are integers. While the above results do not address the existence of limit cycles, they provide the period of candidate limit cycles to within an integer multiple. Equivalently, they provide the location of possible spectral peaks in the output periodogram. As an example, consider the double-loop modulator with dc input $X = 0.5$, and pole values $p_1 = 1.02$ and $p_2 = 1$. The smallest integer N that satisfies (3) is $N = 49$, with resulting $(N_1,\ N_2) = (37,\ 12)$. It is easy to show via simulations that for the above dc input and pole location and initial conditions $[u(0),\ v(0)] = (0,\ 0)$, the spacing between spectral peaks in the output periodogram indicates periodic behavior with period $2 * 49$. Initial conditions $[u(0), v(0)] = (2,\ 5)$ result in spectral peaks corresponding to periodic behavior with period 49.

B. Case 2: $p_1 > 1$ and $p_2 > 1$

For the case where both p_1 and p_2 are outside the unit circle but close to 1, approximations for the period and the number of $+1$'s and -1's to within an integer multiple of possible unstable

Manuscript received February 14, 1996; revised March 17, 1997. This paper was recommended by Associate Editor I. N. Hajj.

The authors are with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720-1770 USA.

Publisher Item Identifier S 1057-7130(98)00055-X.

[1] We refer to $\Sigma\Delta$ modulators with one or more open-loop poles outside the unit circle as having unstable filter dynamics. Ideal $\Sigma\Delta$ modulators refer to modulators with open-loop poles on the unit circle.

[2] Our simulations indicate that fixing p_1 on the unit circle and moving p_2 outside the unit circle results in a worse tone behavior than fixing p_2 on the unit circle and allowing p_1 to move outside the unit circle. Referring to Fig. 1, we note that p_2 only affects state variable $u(n)$, while p_1 affects both state variables. Thus, intuitively, keeping p_1 farther away from the unit circle is more effective in randomizing state-space trajectories.

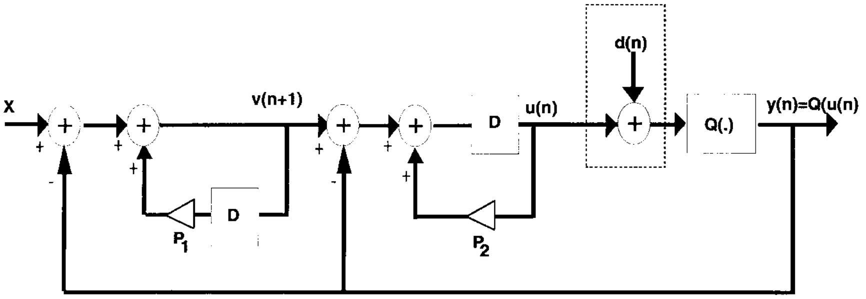


Fig. 1. The double-loop $\Sigma\Delta$ modulator.

limit cycles are given by an equation similar to (3) in Section III-A. In this case, however, N_1 and N_2 in (3) are evaluated by using the approximation $\hat{m}_y \approx p_1 X$. We refer to (3) which uses the approximation $\hat{m}_y \approx p_1 X$ as approximation (3). In the remainder of this subsection, we consider the location of dominant frequency peaks, and relate them to approximation (3).

Extensive simulations of the output periodogram indicate the existence of a high-frequency peak at f_h and a lower frequency peak at f_b:

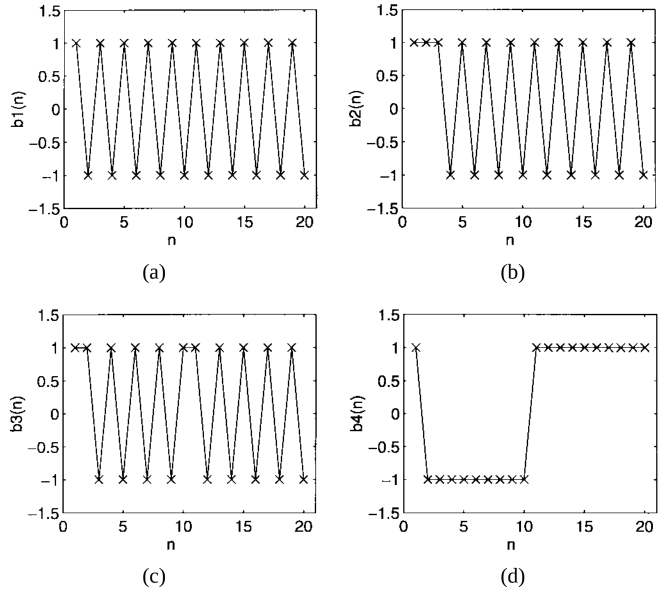
$$f_h \approx (1 - p_1 X) f_s/2 \tag{4}$$

and

$$f_b \approx p_1 X f_s. \tag{5}$$

Furthermore, our simulations indicate that for pole locations close to but outside the unit circle and dc inputs near zero, the spectrum will not exhibit tones at frequency locations below f_b or above f_h.[3] Specifically, we found that inputs less than or equal to 0.031 (0.05) result in spectra whose lowest (highest) frequency peaks are located at $f_b(f_h)$.

In what follows, we present plausibility arguments justifying the above expressions for the predominant spectral peaks.[4] We begin by considering a bit pattern with length 20 and average value zero. For this average value, the bit pattern $b_1(\cdot)$ that yields the highest possible frequency components consists of alternating $+1$'s and -1's, and is shown in Fig. 2(a). Now, if the average value is moved slightly away from zero, to 1/10, the minimum length sequence with average value 1/10 has 11 $+1$'s and 9 -1's, i.e., $N = 20$, $N_1 = 11$, and $N_2 = 9$. One way to obtain this sequence is by shifting one of the -1 bits in $b_1(\cdot)$ to $+1$, as shown in Fig. 2(b). However, our simulations and experiments suggest that the predominant output pattern is one that results in the highest frequency components. This pattern is obtained by interweaving the $+1$'s and -1's in the bit pattern in such a way that they are evenly distributed among each other, thus resulting in high-frequency tone components. An example of such a bit pattern $b_3(\cdot)$ is shown in Fig. 2(c). There are two more $+1$'s than -1's in pattern $b_3(\cdot)$. To accommodate these two extra $+1$'s, we distribute them evenly over the duration of the bit pattern: one extra $+1$ occurs right at the beginning, and the other one occurs in the middle of the sequence. Each of these extra $+1$'s is said to result in "bit shifting." More generally, since we have to accommodate $(N_1 - N_2)$ extra $+1$'s among a total of N bits, we are likely to get a "pseudoperiodic" sequence with "period" $N/(N_1 - N_2)$, resulting in a tone at $[(N_1 - N_2)/N]f_s$, as shown in (5). As an example, bit pattern $b_3(\cdot)$ looks periodic with period $N/(N_1 - N_2) = 10$, and has a spectral component at $f_b = f_s/10$.



(a) (b) (c) (d)

Fig. 2. (a) Pattern $b_1(\cdot)$ with average value 0. (b) Pattern $b_2(\cdot)$ with average value 0.1. (c) Pattern $b_3(\cdot)$ with average value 0.1, and evenly distributed $+1$'s and -1's. (d) Pattern $b_3(\cdot)$; $b_4(n) = b_1(n) \times b_3(n)$.

Another consequence of the above bit shifting is that the tone at $f_s/2$, present in the spectrum of $b_1(\cdot)$, moves to a new frequency location. Note that $b_3(\cdot)$ is obtained by multiplying $b_1(\cdot)$ with the sequence $b_4(\cdot)$, shown in Fig. 2(d), which contains a strong frequency component at $[(N_1 - N_2)/2N]f_s$. In the frequency domain, the spectrum of $b_1(\cdot)$ is convolved with the spectrum of $b_4(\cdot)$, resulting in the tone at $f_s/2$ getting modulated by the tone at $[(N_1 - N_2)/2N]f_s$ to $f_h = [1 - (N_1 - N_2)/N](f_s/2)$.

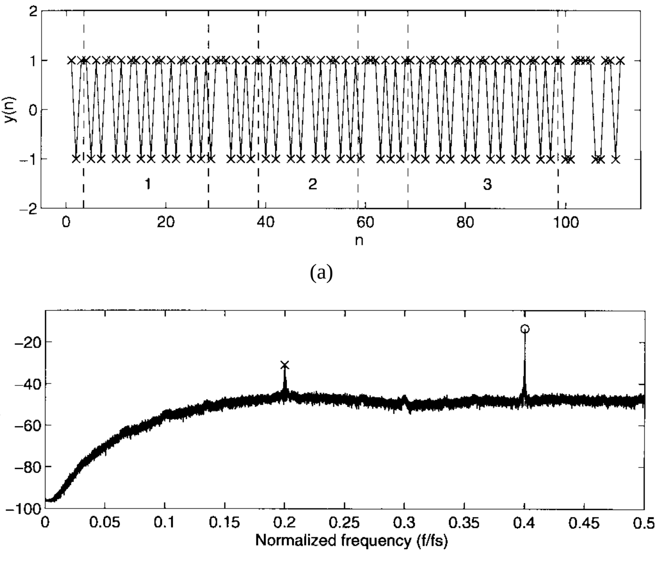
Consider the double-loop modulator with dc input $X = 0.198$ and poles $(p_1, p_2) = (1.01, 1.03)$. The smallest integer N that satisfies approximation (3) is $N = 5$, so that $(N_1, N_2) = (3, 2)$, and the resulting minimum length pattern b that yields the highest possible frequency components is $b = (1, -1, 1, -1, 1)$. Fig. 3(a) shows the output bit sequence for 120 time steps. The regions separated by vertical dashed lines and labeled 1, 2, and 3 in Fig. 3(a) contain portions of the output bit sequence that coincide with pattern b. More generally, the percentage of occurrence of pattern b in the output bit stream is approximately 65%. The dominance of pattern b in the output bit stream is best illustrated by considering the output periodogram. Fig. 3(b) shows the output periodogram obtained by averaging over 20 discrete Fourier transforms (DFT), each 60 000 points long. The output periodogram contains dominant tones at $f_b = \frac{1}{5} f_s$ and $f_h = \frac{2}{5} f_s$, denoted by $\times$ and $\circ$ in Fig. 3(b), respectively. These tones are in agreement with the dominant frequencies of pattern b.

C. General Trends

In this section, we show via simulations that poles slightly outside the unit circle are more advantageous in both suppressing *low-*

[3] We note that this behavior does not hold when the open-loop poles are on or inside the unit circle [7].

[4] Similar arguments are presented in [6] for the modulator with both poles on the unit circle.



(a)

(b)

Fig. 3. $X = 0.198$, $(p_1, p_2) = (1.01, 1.03)$. (a) Output bit sequence (time domain). (b) Output periodogram; ×: frequency location f_b, o: frequency location f_h.

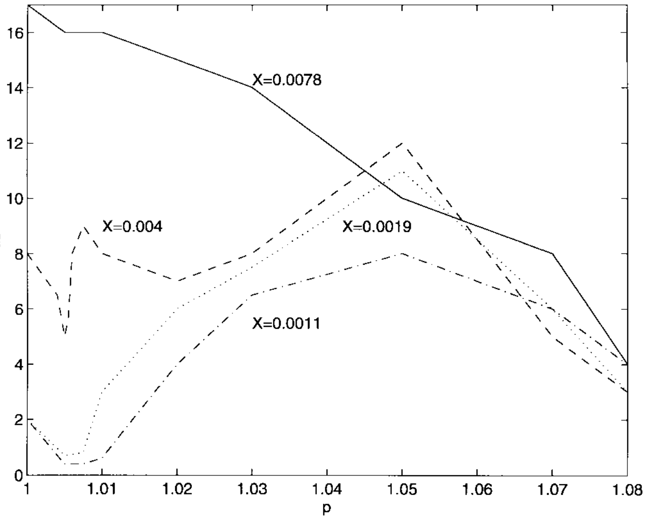


Fig. 4. Relative amplitude of the tone at f_b as a function of open-loop poles, $p = p_1 = p_2$. -: $X = 0.0078$; - - -: $X = 0.004$; $\cdots$: $X = 0.0019$; -.-.: $X = 0.0011$.

frequency tones and maintaining SNR than large pole values outside the unit circle. As pole moduli outside the unit circle are moved toward the unit circle, the noise-shaping properties are improved, thus reducing the level of low-frequency peaks. Fig. 4 shows the relative amplitude of the tone at f_b as a function of pole location $p = p_1 = p_2$ for dc input values $X = 0.0011$, 0.0019, 0.004, and 0.0078. These values are chosen to be the largest dc inputs for which frequency location f_b occurs at the edge of the baseband for the OSR of 512, 256, 128, and 64, respectively. Note that, from (5), f_b falls in the signal baseband for dc values, $X \leq 1/2p_1$ OSR, where the signal baseband is defined as $B = f_s/2\text{OSR}$.

For $X = 0.0011$, 0.0019, and 0.004 in Fig. 4, the amplitude of the tone at f_b initially increases as p is increased from 1.005 to 1.05, and then decreases as p is further increased beyond 1.05. On the other hand, the allowable input dynamic range and SNR are reduced monotonically as p increases from 1.005 to 1.08. For instance, for pole values greater than 1.08, the maximum allowable input that results in bounded internal behavior is less than $X = 0.2$ [5].

Since, for the OSR of 256, $X = 0.0019$ is the largest dc input for which f_b falls into the baseband, and since, from Section III-B, for all $X \leq 0.0019$ f_b is the location of the lowest frequency tone, the shape of the curves in Fig. 4 suggests that at the OSR of 256, the "best" choice for pole values is around 1.005. However, from Fig. 4, $p_1 = p_2 = 1.005$ is not a good choice when the OSR is 64. Indeed, as observed by Dunn and Sandler [8], one might have to choose much larger pole values such as 1.08 in order to reduce the level of baseband tones. Such large values of poles also result in a significant drop in SNR [5], [8].

IV. Performance of Dithering Versus Pole Placement

Dithering is commonly used to improve the subjective quality of quantized signals [2]. Since in Section III-C we concluded that for large OSR's such as 256, poles slightly outside the unit circle, e.g., $p_1 = p_2 = 1.005$, perform "best" in terms of low-frequency tone removal and SNR, in this section, we compare the performance of the above pole placement technique with that of dithering[5] for the OSR of 256. In comparing these two approaches, we will consider both peak signal-to-noise ratio (PSNR) and low-frequency tone behavior.

We consider two commonly used dither characteristics, namely, 1- and 8-bit quantized random dither signals, each with peak-to-peak amplitudes of 1 and uniform probability distribution functions. The level of baseband tones that results in an audible distortion is assessed via experiments done on a Crystal Semiconductors evaluation board, CS4303, along with a set of speakers with bandwidth up to 15 kHz and a wide-band spectrum analyzer. Our experiments indicate that output periodograms containing baseband tones that are 4 dB above the noise floor[6] result in *audible* tones, while those containing tones below 4 dB are not discernible.

Based on our results in Section III-C, dc inputs below $X = 1.9 * 10^{-3}$ may result in baseband tones at frequency locations less than or equal to $f_b = p_1 X f_s$ for an OSR of 256. Furthermore, the double-loop $\Sigma\Delta$ modulator with an OSR of 256 approximately yields 16 bits of resolution [9]. Thus, the smallest dc input that can be differentiated from zero input is approximately $1.6 * 10^{-5}$. Consequently, we evaluate tonal behavior for inputs in the range $R1 = [1.6 * 10^{-5},\ 1.9 * 10^{-3}]$. For the sake of completeness, we also show results for larger inputs in the range $R2 = [0.1,\ 0.8]$.

Fig. 5 shows PSNR values and baseband tonal content[7] for four classes of modulators and 20 dc input values in the range $R1 = [1.6 * 10^{-5},\ 1.9 * 10^{-3}]$ and $R2 = [0.1, 0.8]$. As seen, the performance of the modulator with unstable filter dynamics with poles at $(p_1, p_2) = (1.005, 1.005)$ is comparable to that of additive 8-bit dither, both in terms of PSNR performance and in terms of tonal content. Here, our dither results are comparable to those obtained in [8]. Even though the unstable modulator's tones are up to 1 dB above noise floor for certain dc inputs, in practice, these tones are not audible since they are still below 4 dB.

Throughout this paper, we have emphasized the case $p_1 = p_2$ over $p_1 \neq p_2$. An interesting question is whether there are other

[5] In all that follows, the double-loop modulator with open-loop poles outside the unit circle corresponds to the block diagram shown in Fig. 1 where signal $d(\cdot)$ is identically zero, while the dithered double-loop modulator corresponds to the block diagram shown in Fig. 1 with $p_1 = p_2 = 1$ and $d(\cdot)$ a given dither signal.

[6] In all that follows, the noise floor refers to the flat portion of the periodogram.

[7] Each PSNR value is obtained by averaging over 20 estimates of the SNR using sinusoidal inputs with frequency at half the baseband. The tone behavior is assessed by averaging over 20 DFT's, each 65 536 points long, of the output bit stream after an initial transient.

	X $(*10^{-4})$											PSNR (dB)
	0.16	1	3	5	7	9	11	13	15	17	19	
$P_1 = P_2 = 1.005$	□	□	▨	▨	□	▥	▥	▥	▨	▥	▥	98
$P_1 = 1.02$, $P_2 = 1.005$	□	▒	▒	▒	▒	▒	▒	▒	■	▒	▒	91
Additive 1-bit dither	□	□	□	□	□	□	□	□	□	□	□	92
Additive 8-bits dither	□	□	□	□	□	□	□	□	□	□	□	96

(a)

	X							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$P_1 = P_2 = 1.005$	□	□	□	□	□	□	□	□
$P_1 = 1.02$, $P_2 = 1.005$	□	▒	□	□	□	□	▦	□
Additive 1-bit dither	□	□	□	□	□	□	□	□
Additive 8-bits dither	□	□	□	□	□	□	□	□

□ No tones

▨ Tones < 0.5 dB above the noise floor.

▥ Tones < 1 dB above the noise floor

▒ Tones <= 2 dB above the noise floor

▦ Tones < 4 dB above the noise floor

■ Tones >= 4 dB above the noise floor

(b)

Fig. 5. Baseband tonal content and PSNR values for various dc input values, pole locations, and additive dither signals.

combinations of p_1 and p_2 besides $p_1 = p_2 = 1.005$ that result in superior PSNR and tone performance. One possibility is to keep p_2 close to but outside the unit circle in order to preserve the PSNR advantage, e.g., $p_2 = 1.005$, and move p_1 further away from the unit circle to preserve the randomness in state space trajectories,[8] e.g., $p_1 = 1.02$. However, Fig. 5 clearly shows that $p_1 = p_2 = 1.005$ outperforms $(p_1,\ p_2) = (1.02,\ 1.005)$ in both PSNR and tonal characteristics. In fact, the PSNR performance of the modulator with $(p_1,\ p_2) = (1.02,\ 1.005)$ is comparable to that of the modulator with additive 1-bit dither, while its tonal behavior is slightly worse. Specifically, the modulator with 1-bit dither has no audible tones, while the modulator with $(p_1,\ p_2) = (1.02,\ 1.005)$ has an audible tone for the dc input of $X = 0.0015$.

In terms of hardware implementation for A/D conversion applications, extra analog hardware is needed to implement a random noise source for the dithered modulator. Pole placement outside the unit circle, however, may be achieved by either appropriately designing capacitor ratios in the integrators or by applying positive feedback in the operational amplifiers [10]. The complexity in generating random dither signals for A/D applications may be reduced by quantizing the dither signal to 1 bit [8]. However, this results in a large loss in SNR performance. As seen from Fig. 5, the modulator with 1-bit additive dither results in 6 and 4 dB lower PSNR than modulators with $p_1 = p_2 = 1.005$ and 8-bit additive dither, respectively. Thus, for large OSR's, the $\Sigma\Delta$ modulator with poles slightly outside the unit circle is an attractive alternative to the dithered modulator when hardware complexity is desired to be minimal.

Acknowledgment

The authors would like to thank Prof. B. Boser, Dr. M. Varghese, and Dr. D. Senderowicz for valuable discussions and comments.

References

[1] R. M. Gray, "Spectrial analysis of quantization noise in a single loop sigma delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588–599, June 1989.

[2] S. Norsworthy and D. A. Rich, "Idle channel tones and dithering in $\Sigma\Delta$ modulators," presented at the *9th AES Conv.*, Oct. 1993.

[3] M. Motamed, A. Zakhor, and S. Sanders, "Tones, saturation and SNR in double loop $\Sigma\Delta$ modulators," in *Proc. ISCAS*, 1993, pp. 1345–1348.

[4] S. Hein, "Exploiting chaos to suppress spurious tones in general double-loop $\Sigma\Delta$ modulators," *IEEE Trans. Circuit Syst. II*, vol. 40, pp. 651–659, Oct. 1993.

[5] M. Motamed, S. Sanders, and A. Zakhor, "The double loop sigma delta modulator with unstable filter dynamics: Stability analysis and tone behavior," *IEEE Trans. Circuit Syst. II*, vol. 43, pp. 549–559, Aug. 1996.

[6] R. C. Ledzius and J. Irwin, "The basis and architecture for the reduction of tones in a sigma–delta DAC," *IEEE Trans. Circuit Syst. II*, vol. 40, pp. 429–439, July 1993.

[7] M. Motamed, "The double loop sigma–delta modulator with unstable filter dynamics: Stability analysis and tone behavior," Ph.D. dissertation, Univ. California, Berkeley, Apr. 1996.

[8] C. Dunn and M. Sandler, "Linearizing sigma–delta modulators using dither and chaos," in *Proc. ISCAS*, 1995, pp. 625–628.

[9] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," *Oversampling Delta–Sigma Data Converters. Theory, Design and Simulation*, J. C. Candy and G. C. Temes, Ed. New York: IEEE, 1991.

[10] D. Senderowicz, M. Motamed, A. Zakhor, and D. Clementi, "Using chaos to eliminate idle tones in sigma delta converters," in preparation.

[8] As mentioned in Section III, this results in a better tone behavior than keeping p_1 close to the unit circle and moving p_2 further out.