One-Bit Dithering in Delta-Sigma Modulator-Based D/A Conversion

Ian Galton

Department of Electrical and Computer Engineering
University of California
Irvine, CA 92717

Abstract—Although delta-sigma modulators are widely used in oversampling D/A converters, their requantization error can contain discrete tones that are objectionable in high-fidelity audio applications. This paper investigates the asymptotic second-order statistics and ergodicity of the requantization error and shows that such tones can be eliminated in a class of delta-sigma modulators with least significant bit dithering.

I. Introduction

Delta-sigma ($\Delta\Sigma$) modulator-based D/A converters are well represented in the consumer audio electronics market and show promise for achieving greater than 20 bits of accuracy. However, they suffer from a problem not encountered in their conventional counterparts: limit cycles within the $\Delta\Sigma$ modulators can cause tones of significant power to appear in the requantization error [1], [2]. As the ear is particularly adept at discerning such tones, the behavior is objectionable in high fidelity audio applications.

It has been observed that adding a one-bit random *dither sequence* to the least significant bit of the $\Delta\Sigma$ modulator input can reduce the unwanted tones in the requantization error [1]. However, this observation has not been previously backed up by theory. This paper presents a theoretical basis for the behavior. It shows that limit cycles can be completely avoided for a large class of $\Delta\Sigma$ modulators with a properly chosen random bit sequence. The result has practical significance because it implies that for the relatively small price of a feedback shift register circuit to generate a pseudo-random bit sequence and an extra input bit in the $\Delta\Sigma$ modulator circuitry, the requantization error tones can be essentially eliminated.

The remainder of the paper consists of two main sections and an appendix. Section II describes a generic $\Delta\Sigma$ modulator architecture of which many of the known $\Delta\Sigma$ modulators are special cases, and Section III derives various of its properties. These properties are inherited by any $\Delta\Sigma$ modulator that fits the paradigm of the generic $\Delta\Sigma$ modulator. The appendix presents various lemmas that support the results of Section III.

II. A Generic Delta-Sigma Modulator

Many of the published $\Delta\Sigma$ modulator architectures are special cases of the generic $\Delta\Sigma$ modulator shown in Figure 1 [3]. The system consists of a linear time invariant (LTI) digital system, $\mathbf{T}(z)$, followed by a bank of requantizers followed by another LTI digital system, $\mathbf{U}(z)$. A feedback path joins the output of the requantizer bank to the input of $\mathbf{T}(z)$. In an actual circuit, coarse D/A converters would either follow the requantizers directly, or be placed further along in the processing chain. Provided no additional requantization is performed the position of the D/A converters does not affect the analysis.

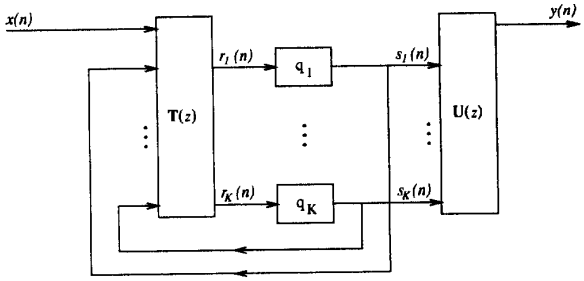

Figure 1: A generic delta-sigma modulator architecture.

The matrix transfer function $\mathbf{T}(z)$ will be written as:

$$\mathbf{T}(z) = \begin{bmatrix} F_1(z) & G_{1,1}(z) & \dots & G_{1,K}(z) \\ \vdots & \vdots & \ddots & \vdots \\ F_K(z) & G_{K,1}(z) & \dots & G_{K,K}(z) \end{bmatrix}$$

and the following two conditions will be assumed:

Condition 1: The requantizers do not overload.

Condition 2: For each j, k, the impulse response $g_{j,k}(n)$ (i.e., the inverse z-transform of $G_{j,k}(z)$) only takes on values that are integer multiples of Δ_j/Δ_k.

As discussed in [3], these conditions are sufficiently mild that a large class of the published $\Delta\Sigma$ modulators, including most of the multistage architectures, are special cases of the generic $\Delta\Sigma$ modulator.

By definition, the input sequence to any D/A converter only takes on a finite number of discrete values and it is impractical not to place the same restriction on the input to each requantizer within the $\Delta\Sigma$ modulator. Consequently, the input sequence and the sequences applied to each requantizer must always be multiples of some *minimum step-size*, δ, and the impulse responses

0-7803-1254-6/93$03.00 © 1993 IEEE

$f_k(n)$ (i.e., the inverse z-transform of $F_k(z)$) and $g_{j,k}(n)$ must be integer-valued. For example, in a $\Delta\Sigma$ modulator implementation wherein multi-bit binary numbers are used to represent numerical values (prior to the coarse D/A conversion), δ would correspond to the value of the least significant bit.

The output of the $\Delta\Sigma$ modulator can be considered the sum of two components: a *signal component* corresponding to the input sequence and a *requantization error sequence* arising from the requantization process. Without loss of generality, each requantizer can be considered a device that outputs the sum of its input sequence and a *requantization noise sequence*. The requantization error sequence is therefore the component of the output sequence corresponding to the requantization noise sequences from all the requantizers. For example, referring to Figure 1, the requantization noise sequence introduced by the k^{th} requantizer must be $\epsilon_k(n) = s_k(n) - r_k(n)$. As shown in [3], the requantization error sequence is equal to the requantization noise sequences filtered by

$$\mathbf{N}(z) = \mathbf{U}(z)\big(\mathbf{I} - \mathbf{G}(z)\big)^{-1},$$

where $\mathbf{G}(z)$ equals $\mathbf{T}(z)$ with the first column deleted. The signal component is equal to the input sequence filtered by

$$S(z) = \mathbf{U}(z)\big(\mathbf{I} - \mathbf{G}(z)\big)^{-1}\mathbf{F}(z),$$

where $\mathbf{F}(z)$ is the first column of $\mathbf{T}(z)$.

The results of this paper correspond to requantizers that perform uniform mid-rise quantization with stepsizes $\Delta_1, \ldots, \Delta_K$, respectively, where each Δ_k is an even-integer multiple of δ. Such is the case when two's complement arithmetic is used, and each requantizer truncates its input and adds $\Delta_k/2$ to the result. Similar results can be obtained for other types of uniform requantizers.

III. Requantization Noise Statistics

Define the *actual input sequence*, $x(n)$, to be the sum of the *desired input sequence*, $x_d(n)$, plus a *dither sequence*, $d(n)$. The desired input sequence is the sequence that is to be converted into an analog waveform, and the dither sequence is a sequence of independent identically distributed (iid) random variables. Because both the desired input and dither sequences are restricted to amplitudes that are multiples of δ, each member of either sequence must have a distribution consisting of point masses at integer multiples of δ.

Let n_0 be the time that the system is "turned on". As shown in [3], for each $n \geq n_0$, the requantization noise can be written as

$$\epsilon_k(n) = \frac{\Delta_k}{2} - \Delta_k\left\langle \frac{1}{\Delta_k}\left[\alpha_k(n) + \sum_{m=0}^{n-n_0} f_k(m)x(n-m)\right]\right\rangle, \quad (1)$$

where $f_j(n)$ is the inverse z-transform of $F_j(z)$ and $\alpha_k(n)$ is a deterministic function of $\mathbf{T}(z)$ and n that does not depend upon the input sequence. The angle brackets denote the fractional part operator. For all $n < n_0$, $\epsilon_k(n)$ will be taken to be zero.

The following two theorems present sufficient conditions for the requantization noise to be asymptotically white, independent of $x_d(n)$, and uniformly distributed as a function of run-time. The results are sufficient to ensure that limit cycles are eliminated.

Theorem 1: Suppose the probability distribution of the dither sequence is non-zero on at least two consecutive multiples of δ and that $\{mf_k(n) \bmod \frac{\Delta_k}{\delta}\}$ does not converge to zero as $n \to \infty$ for any $m = 1, \ldots, \frac{\Delta_k}{\delta} - 1$. Then, as $n_0 \to -\infty$, $\epsilon_k(n)$ converges in distribution to a random sequence $\epsilon'_k(n)$ that is independent of $x_d(n)$ and is uniformly distributed on the set of amplitude points $\{i\delta : i = -\frac{\Delta_k}{2\delta} + 1, -\frac{\Delta_k}{2\delta} + 2, \ldots, \frac{\Delta_k}{2\delta}\}$.

Proof: For each k, (1) has the form $\epsilon_k(n) = \frac{\Delta_k}{2} - \Delta_k U_{n-n_0}$ where U_{n-n_0} is a random variable satisfying the hypothesis of Lemma 2 (see Appendix) with

$$\mu_{n-n_0} = \frac{1}{\Delta_k}\left[\alpha_k(n) + \sum_{m=0}^{n-n_0} f_k(m)x_d(n-m)\right],$$

$b_i = f_k(i)\delta/\Delta_k$, $\eta_i = d(n-i)/\delta$, and $N_0 = \Delta_k/\delta$.

If $\{mf_k(n) \bmod \frac{\Delta_k}{\delta}\}$ does not converge to zero then $\langle mb_i \rangle$ must not converge to zero either. Lemma 2(i) therefore implies that as $n_0 \to -\infty$, U_{n-n_0} converges in distribution to a random sequence U_n that is uniformly distributed on the set of amplitude points $\{i/N_0 : i = 0, \ldots, \frac{\Delta_k}{\delta} - 1\}$. Hence, $\epsilon_k(n)$ converges in distribution to a random sequence $\epsilon'_k(n)$ that is uniformly distributed on $\{i\delta : i = -\frac{\Delta_k}{2\delta} + 1, \ldots, \frac{\Delta_k}{2\delta}\}$. Moreover, since the convergence is uniform with respect to $x_d(n)$ and since the distribution of $\epsilon'_k(n)$ is not conditioned on $x_d(n)$, $\epsilon'_k(n)$ must be independent of $x_d(n)$.

■

Theorem 2: Suppose in addition to the hypothesis of Theorem 1 that for $p \neq 0$, $\{[m_0 f_k(n) + m_1 f_k(n+p)] \bmod \frac{\Delta_k}{\delta}\}$ does not converge to zero for for any pair of integers, m_0 and m_1, not both zero, such that $0 \leq m_0, m_1 \leq \frac{\Delta_k}{\delta} - 1$. Then, $\epsilon'_k(n)$ and $\epsilon'_k(n+p)$ are independent.

Proof: The proof is similar to that of Theorem 1 except that it relies upon Lemma 2(ii) instead of Lemma 2(i).

■

One of the simplest dither sequences that satisfies the hypotheses of Theorems 1 and 2 takes on values of δ and 0 with fixed probabilities. This corresponds to

dithering the least significant bit of the input to the $\Delta\Sigma$ modulator.

For particular $\Delta\Sigma$ modulator architectures, it is easy to determine whether each $f_k(n)$ satisfies the hypotheses of Theorems 1 and 2. For example, the first-order $\Delta\Sigma$ modulator satisfies the hypothesis of Theorem 1 but not Theorem 2 while the single-loop second-order $\Delta\Sigma$ modulator [1] satisfies the hypotheses of both theorems. Indeed, many of the common $\Delta\Sigma$ modulator architectures satisfy the hypotheses of both theorems.

In accordance with the usual definitions, take the mean and autocovariance of the quantization noise from the k^{th} requantizer to be

$$M_{\epsilon_k} = \lim_{n_0 \to -\infty} \mathrm{E}[\epsilon_k(n)], \tag{2}$$

and

$$C_{\epsilon_k\epsilon_k}(p) = \lim_{n_0 \to -\infty} \mathrm{E}[(\epsilon_k(n) - M_{\epsilon_k})(\epsilon_k(n+p) - M_{\epsilon_k})], \tag{3}$$

respectively.

It follows directly from Theorem 1 that $M_{\epsilon_k} = \frac{\delta}{2}$. Moreover, Theorem 2 implies that each requantization noise sequence is asymptotically white with autocovariance

$$C_{\epsilon_k\epsilon_k}(p) = \begin{cases} \frac{\Delta_k^2}{12} + \frac{\delta^2}{6}, & \text{if } p = 0; \\ 0, & \text{otherwise.} \end{cases}$$

In most of the common $\Delta\Sigma$ modulators, $\mathbf{N}(z)$ has only one non-zero element. In such cases, if $x_d(n)$ is a wide-sense stationary or quasi-stationary sequence, it follows that the power spectral density (PSD) of the overall $\Delta\Sigma$ modulator output, neglecting DC terms, can be written as

$$S_{yy}(e^{j\omega}) = S_{xx}(e^{j\omega})\left|S(e^{j\omega})\right|^2 + \left[\frac{\Delta_k^2}{12} + \frac{\delta^2}{6}\right]\left|N_\sigma(e^{j\omega})\right|^2,$$

where $S_{xx}(e^{j\omega})$ is the PSD of $x(n)$, and $N_\sigma(e^{j\omega})$ is the non-zero element of $\mathbf{N}(z)$.

For the results presented so far to be of practical value, it is necessary that the statistical averages (2) and (3) equal the corresponding time averages and that there is no average time correlation between the requantization noise and the input sequence. These properties are asserted by the following theorem.

Theorem 3: If the hypothesis of Theorem 1 is satisfied,

$$\frac{1}{N}\sum_{n=0}^{N-1} \epsilon_k(n) \to M_{\epsilon_k} \tag{4}$$

and

$$\frac{1}{N}\sum_{n=0}^{N-1} x_d(n)(\epsilon_k(n+p) - M_{\epsilon_k}) \to 0 \tag{5}$$

in probability as $N \to \infty$. Moreover, if the hypothesis of Theorem 2 is satisfied,

$$\frac{1}{N}\sum_{n=0}^{N-1}(\epsilon_k(n) - M_{\epsilon_k})(\epsilon_k(n+p) - M_{\epsilon_k}) \to C_{\epsilon_k\epsilon_k}(p) \tag{6}$$

in probability as $N \to \infty$.

Proof: Because the proofs of (4), (5), and (6) are similar, only the proof of (6) will be provided.

Let $X_n = (\epsilon_k(n) - M_{\epsilon_k})(\epsilon_k(n+p) - M_{\epsilon_k}) - C_{\epsilon_k\epsilon_k}(p)$. Then, it is sufficient to show that

$$\frac{1}{N}\sum_{n=0}^{N-1} X_n \to 0$$

holds in probability. From the proof of Theorem 2, it follows that for every integer $j > n_0$,

$$\mathrm{E}_{\{d(i):i\geq j\}}[X_n] \to 0$$

uniformly with respect to the variables $\{\eta_{n_0}, \ldots, \eta_{j-1}\}$ and $\{x_d(n)\}$. The result follows from Lemma 3. ■

Another important and largely unanswered question relates to the rate of the asymptotic convergence. The results above characterize the long-term average behavior of the quantization noise, but do not give insight into its short-term behavior. Provided the asymptotic convergence is rapid enough compared to the input oversampling ratio, the short-term behavior is typically not a significant concern. In general, the results above indicate that the rate of the asymptotic convergence tends to be a strong function of the specific $\Delta\Sigma$ modulator architecture and the $F_k(z)$ functions in particular.

IV. Appendix

This appendix presents three lemmas that provide the basis for the theorems presented above. The first and third lemmas were presented in [3] and [4], respectively, and are repeated here for convenience. The second lemma is believed by the author to be new. In this section, the letter Φ is used to denote the characteristic function of its subscript variables.

Lemma 1: For each $p = 0, 1, \ldots$, let

$$X_p = \mu_p + \sum_{k=0}^{p} b_k\eta_k, \qquad Y_p = \nu_p + \sum_{k=0}^{p} c_k\eta_k,$$

$U_p = \langle X_p \rangle$, and $V_p = \langle Y_p \rangle$, where $\{\eta_k\}$ is an iid sequence of random variables, $\{\mu_p\}$ and $\{\nu_p\}$ are any sequences of random variables that are independent of each η_k, and $\{b_k\}$ and $\{c_k\}$ are any deterministic sequences. Then

$$\Phi_{X_p,Y_p}(t_0,t_1) = \Phi_{\mu_p,\nu_p}(t_0,t_1)\prod_{k=0}^{p}\Phi_\eta(t_0b_k + t_1c_k).$$

Furthermore,

$$\Phi_{U_p,V_p}(2\pi m_0, 2\pi m_1) = \Phi_{X_p,Y_p}(2\pi m_0, 2\pi m_1),$$

for every pair of integers m_0, m_1.

Lemma 2: Let N_0 and N_1 be positive integers. In the hypothesis of Lemma 1 suppose for every k that μ_k and b_k only assume integer multiples of $1/N_0$, that ν_k and c_k only assume integer multiples of $1/N_1$, and that η_k is integer-valued with a probability distribution that is non-zero on at least two consecutive integers.

(i) If $\langle mb_k \rangle$ does not converge to zero for any $m = 1, \ldots, N_0 - 1$, then U_p converges in distribution to a random variable U that is uniformly distributed on the set $\{i/N_0 : i = 0, \ldots, N_0 - 1\}$.

(ii) If $\langle m_0 b_k + m_1 c_k \rangle$ does not converge to zero for any $m_0 = 0, \ldots, N_0 - 1$ and $m_1 = 0, \ldots, N_1 - 1$ except $m_0 = m_1 = 0$ then U_p and V_p converge in distribution to independent random variables U and V, respectively.

Proof: To show that U_p and V_p converge in distribution to U and V, respectively, it is sufficient to show that $\Phi_{U_p,V_p}(t_0, t_1)$ converges to a characteristic function $\Phi_{U,V}(t_0, t_1)$ as $p \to \infty$. By definition, (U_p, V_p) only assumes values from the set $\{(m_0/N_0, m_1/N_1) : m_0 = 0, \ldots, N_0 - 1; m_1 = 0, \ldots, N_1 - 1\}$. Hence, the samples $\{\Phi_{U_p,V_p}(2\pi m_0, 2\pi m_1) : m_0 = 0, \ldots, N_0 - 1; m_1 = 0, \ldots, N_1 - 1\}$ uniquely determine $\Phi_{U_p,V_p}(t_0, t_1)$. It is therefore sufficient to show that $\Phi_{U_p,V_p}(t_0, t_1)$ converges to $\Phi_{U,V}(t_0, t_1)$ at these sample points.

The characteristic function common to each η_k is

$$\Phi_\eta(t) = \sum_{n=-\infty}^{\infty} p_\eta(n) e^{jtn},$$

where $p_\eta(n)$ is the probability distribution of each η_k. By hypothesis, there must be some pair of consecutive integers k_0 and k_1 such that $p_\eta(k_0) \neq 0$ and $p_\eta(k_1) \neq 0$. Without loss of generality, assume that $p(k_0) \leq p(k_1)$. Applying the triangle inequality gives

$$\begin{aligned} \left|\Phi_\eta(t)\right| &\leq \left| p_\eta(k_0)\left[e^{jtk_0} + e^{jtk_1}\right]\right| \\ &\quad + \left|\left[p_\eta(k_1) - p_\eta(k_0)\right] e^{jtk_1}\right| + \sum_{n \neq k_0, k_1} \left|p_\eta(n) e^{jtn}\right| \\ &= p_\eta(k_0)\left[2\left|\cos(t/2)\right| - 1\right] + \sum_{n \neq k_0} p_\eta(n). \end{aligned}$$

Thus, $\left|\Phi_\eta(t)\right| < 1$ for all t not equal to integer multiples of 2π.

From Lemma 1, it follows that

$$\lim_{p \to \infty} \Phi_{U_p,V_p}(2\pi m_0, 2\pi m_1) = 0,$$

provided $\langle m_0 b_k + m_1 c_k \rangle$ does not converge to zero. Furthermore, since all characteristic functions equal one at the origin, $\Phi_{U_p,V_p}(0,0) = 1$ for all p.

Because $\Phi_{U_p}(t) = \Phi_{U_p,V_p}(t, 0)$, it follows from the hypothesis of (i) that $\lim_{p \to \infty} \Phi_{U_p}(2\pi m) = \Phi_U(2\pi m)$ where

$$\Phi_U(2\pi m) = \begin{cases} 1, & \text{if } m = 0; \\ 0, & \text{if } m = 1, \ldots, N_0 - 1. \end{cases}$$

These samples uniquely specify the characteristic function

$$\Phi_U(t) = \frac{1}{N_0} e^{\frac{jt}{2}(1 - \frac{1}{N_0})} \frac{\sin\left(\frac{t}{2}\right)}{\sin\left(\frac{t}{2N_0}\right)},$$

which corresponds to a random variable, U, that is uniformly distributed on $\{i/N_0 : i = 0, \ldots, N_0 - 1\}$.

Similarly, it follows from the hypothesis of (ii) that

$$\Phi_{U,V}(t_0, t_1) = \frac{e^{\frac{jt_0}{2}(1 - \frac{1}{N_0})} e^{\frac{jt_1}{2}(1 - \frac{1}{N_1})}}{N_0 N_1} \frac{\sin\left(\frac{t_0}{2}\right)}{\sin\left(\frac{t_0}{2N_0}\right)} \frac{\sin\left(\frac{t_1}{2}\right)}{\sin\left(\frac{t_1}{2N_1}\right)},$$

corresponding to a random vector, (U, V), that is uniformly distributed on $\{(i_0/N_0, i_1/N_1) : i_0 = 0, \ldots, N_0 - 1, i_1 = 0, \ldots, N_1 - 1\}$. Hence, U and V are independent. ∎

Lemma 3: For each $k = 0, 1, \ldots$, let X_k be a deterministic function of the random sequences $\{\chi_0, \ldots, \chi_k\}$ and $\{\eta_0, \ldots, \eta_k\}$, where the η_n are independent random variables that are independent of the χ_n. Suppose that the distribution of each X_k has its support restricted to $[-\beta, \beta]$ where $\beta \in \mathbf{R}$, and that for each non-negative integer j, as $k \to \infty$

$$\mathop{\mathrm{E}}_{\{\eta_n : n > j\}}(X_k) \to 0$$

uniformly with respect to the variables $\{\eta_0, \ldots, \eta_j\}$ and $\{\chi_0, \chi_1, \ldots\}$. Then

$$\frac{1}{N} \sum_{n=0}^{N-1} X_n \to 0$$

in probability as $N \to \infty$.

References

1. J. C. Candy, G. C. Temes, "Oversampling Methods for A/D and D/A Conversion," *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, New York: IEEE Press, pp. 1-25, 1992.
2. R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1220-1244, Nov. 1990.
3. I. Galton, "Granular quantization noise in a class of delta-sigma modulators," Submitted to *IEEE Trans. Inform. Theory*, Mar. 1992
4. I. Galton, "Granular quantization noise in the first-order delta-sigma modulator," Submitted to *IEEE Trans. Inform. Theory*, Nov. 1991.