Midwest Symposium on Circuits and Systems 2000
Lansing, MI, August 2000

Joint Shaping of Quantization and Hardware-Mismatch Errors in a Multibit Delta-Sigma DAC

Dan P. Scholnik and Jeffrey O. Coleman
scholnik@nrl.navy.mil *jeffc@alum.mit.edu*
Naval Research Laboratory, Washington, DC

Abstract— **A recent sigma-delta modulator architecture incorporates spectral shaping of hardware-mismatch errors through redundancy introduced by embedding the input/output subspace in a higher-dimensional space for loop operation. Here basic design issues for the key vector quantizer and matrix-impulse-response loop filter are explored. A simple design example and simulation illustrate.**

I. INTRODUCTION

Recently introduced DAC architectures [1,2,3,4,5,6,7,8,9] use multiple $\Delta\Sigma$-style shaping loops to move hardware-mismatch errors out of the signal band. The primary application of such DACs is in conventional multibit $\Delta\Sigma$ modulators, which require extremely high in-band precision. As shown in [10], a vector input/output $\Delta\Sigma$ modulator and mismatch-shaping DAC can be combined into a single $\Delta\Sigma$-like vector loop, inviting a joint design. Here we look at the design issues raised by the most-common case: a scalar-input, scalar-output multibit DAC operating on an unquantized (continuous-valued) or more finely quantized input.

Multibit $\Delta\Sigma$ modulators lower oversampling requirements (relative to single-bit modulators) by lowering the total quantization-error level and allowing more aggressive error shaping. The step-size errors inherent in a static DAC, however, can severely limit the precision of the $\Delta\Sigma$ output. Error-shaping DACs appear to be the most promising way to achieve the extreme precision required. Redundancy is exploited in the hardware by dynamic element matching (DEM) [11], in which M subconverters (generally one-bit) are used to implement a multibit DAC by combining (summing) their outputs. Since output values can have several realizations (for example, there are M ways to activate exactly one subconverter), switching between these as time progresses has the effect of "averaging out" the errors due to hardware errors. In the sequel, "elements" and "DAC elements" refer to the subconverters, while "the DAC" and "the $\Delta\Sigma$ DAC" refer to the full system.

II. SYSTEM DESCRIPTION

Figure 1 shows the proposed system, a special case of the vector architecture presented in [10]. Conceptually it is similar to a conventional scalar-input, scalar-output $\Delta\Sigma$ loop, but internally it has vector signals, a vector quantizer, and a matrix noise transfer function (NTF). The resulting redundancy is the key to the shaping of hardware errors in the DAC output. Here,

This work was supported by the AMRF-C program (ONR 31) of the Office of Naval Research.

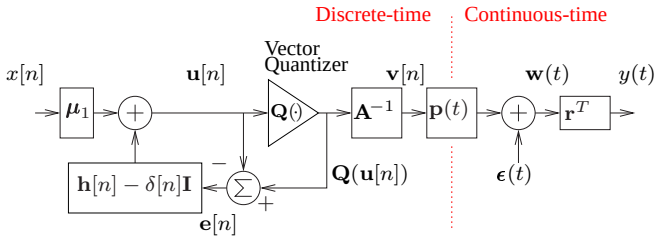

Figure 1: Integrated mismatch-shaping $\Delta\Sigma$ DAC.

$x[n]$ and $y(t)$ are the scalar input and output, internal signals $\mathbf{u}[n]$, $\mathbf{Q}(\mathbf{u}[n])$, $\mathbf{v}[n]$, $\mathbf{e}[n]$, $\mathbf{w}(t)$, and $\boldsymbol{\epsilon}(t)$ are $M \times 1$ vectors, and $\mathbf{A}$, $\mathbf{p}(t)$, and $\mathbf{h}[n]$ are $M \times M$ matrices.

Matrix $\mathbf{A}$ performs a coordinate transformation between "subconverter space", where each vector coordinate corresponds to a single subconverter (for example, signals $\mathbf{v}[n]$ and $\mathbf{w}(t)$), and "signal space", where the first vector coordinate corresponds to the system input/output (signals $\mathbf{u}[n]$ and $\mathbf{e}[n]$). The remaining coordinates are occupied by switching signals that represent the internal redundancy of the system. Usually the first row of $\mathbf{A}$, denoted $\mathbf{r}^T$, is a (scaled) version of $(1,\ldots,1)$ so that output $y(t) = \mathbf{r}^T\mathbf{w}(t)$ is just the (scaled) sum of the individual subconverter outputs. In some architectures [2,12] the remaining rows of $\mathbf{A}$, denoted $\mathbf{S}^T$, can be chosen to reduce hardware complexity. Here we choose simplicity and let $\mathbf{A}$ be an orthogonal matrix (i.e. orthonormal columns), so that $\mathbf{A}^{-1} = \mathbf{A}^T = (\mathbf{r},\ \mathbf{S})$.

The error-feedback structure shown, with loop-filter impulse response $\mathbf{h}[n] - \delta[n]\mathbf{I}$, simplifies the analysis, but any of the usual $\Delta\Sigma$ topologies could be used. Note that for realizability $\mathbf{h}[n] - \delta[n]\mathbf{I}$ must be strictly causal,

$$\begin{aligned}\mathbf{h}[n] &= \mathbf{0}, \quad n < 0 \\ &= \mathbf{I}, \quad n = 0.\end{aligned}$$

A. Hardware Model

The hardware model used is that of [1], which describes a bank of DAC subconverters whose outputs are combined to form a single multibit signal. The subconverters are defined by the matrix pulse response $\mathbf{p}(t)$ and the error signal vector $\boldsymbol{\epsilon}(t)$, which in the case of conventional (unshaped) DAC elements is just a constant offset. Element $\mathbf{p}_{i,j}(t)$ is the unit-sample response of the ith element to the jth input. If all elements have identical output pulses and there is no crosstalk, then $\mathbf{p}(t) = p(t)\mathbf{I}$ for some scalar pulse response $p(t)$. Any

variation among the main-diagonal entries is a result of pulse-height, pulse-shape, and pulse-offset (clock-timing) differences. Nonzero entries outside of the main diagonal model crosstalk between elements. The final DAC output is formed by combining the individual DAC elements in the matrix multiply by (vector projection onto) $\mathbf{r}^T$.

B. Input-Output Relationship

We now proceed to find the transfer characteristic of the system in Fig. 1. The vector quantizer output will be modeled as the sum of the quantizer input and an error term,

$$\mathbf{Q}(\mathbf{u}[n]) = \mathbf{u}[n] + \mathbf{e}[n].$$

Using $*$ for discrete-time convolution, the quantizer input in turn can be written

$$\mathbf{u}[n] = \boldsymbol{\mu}_1 x[n] + (\mathbf{h} * \mathbf{e})[n] - \mathbf{e}[n],$$

with $\boldsymbol{\mu}_1 = (1, 0, 0, \dots)^T$ the first standard unit vector, and substituting this results in

$$\mathbf{Q}(\mathbf{u}[n]) = \boldsymbol{\mu}_1 x[n] + (\mathbf{h} * \mathbf{e})[n].$$

The input to the bank of DAC elements is

$$\begin{aligned}\mathbf{v}[n] &= \mathbf{A}^{-1}\mathbf{Q}(\mathbf{u}[n]) \\ &= \mathbf{r}x[n] + (\mathbf{A}^{-1}\mathbf{h} * \mathbf{e})[n],\end{aligned}$$

recalling that vector $\mathbf{r}$ is the first column of $\mathbf{A}^{-1}$. The quantized output is fed to the bank of DAC subconverters,

$$\begin{aligned}\mathbf{w}(t) &= (\mathbf{p} \star \mathbf{v})(t) + \boldsymbol{\epsilon}(t) \\ &= (\mathbf{p}\mathbf{r} \star x)(t) + (\mathbf{p}\mathbf{A}^{-1} \star \mathbf{h} * \mathbf{e})(t) + \boldsymbol{\epsilon}(t),\end{aligned}$$

where the mixed discrete/continuous-time convolution of (for example) $\mathbf{p}(t)$ and $\mathbf{v}[n]$ is defined as

$$(\mathbf{p} \star \mathbf{v})(t) \triangleq T \sum_k \mathbf{p}(t - kT)\mathbf{v}[k].$$

The system output is $y(t) = \mathbf{r}^T\mathbf{w}(t)$, or

$$y(t) = (\mathbf{r}^T\mathbf{p}\mathbf{r} \star x)(t) + (\mathbf{r}^T\mathbf{p}\mathbf{A}^{-1} \star \mathbf{h} * \mathbf{e})(t) + \mathbf{r}^T\boldsymbol{\epsilon}(t). \quad (1)$$

We can simplify this if we partition $\mathbf{h}[n]$ after the first row/column as

$$\mathbf{h}[n] = \begin{pmatrix} h_x[n] & \mathbf{h}_{x\mathbf{s}}[n] \\ \mathbf{h}_{\mathbf{s}x}[n] & \mathbf{h}_{\mathbf{s}}[n] \end{pmatrix}$$

and $\mathbf{e}[n]$ likewise as

$$\mathbf{e}[n] = \begin{pmatrix} e_x[n] \\ \mathbf{e}_{\mathbf{s}}[n] \end{pmatrix}.$$

We now write the output as

$$\begin{aligned}y(t) &= (p_x \star x)(t) + \mathbf{r}^T\boldsymbol{\epsilon}(t) \\ &\quad + \left[\begin{pmatrix} p_x & \mathbf{p}_{\mathbf{s}} \end{pmatrix} \star \begin{pmatrix} h_x & \mathbf{h}_{x\mathbf{s}} \\ \mathbf{h}_{\mathbf{s}x} & \mathbf{h}_{\mathbf{s}} \end{pmatrix} * \begin{pmatrix} e_x \\ \mathbf{e}_{\mathbf{s}} \end{pmatrix} \right](t), \end{aligned} \quad (2)$$

where $p_x(t) \triangleq \mathbf{r}^T\mathbf{p}(t)\mathbf{r}$ and $\mathbf{p}_{\mathbf{s}}(t) \triangleq \mathbf{r}^T\mathbf{p}(t)\mathbf{S}$. In the $\mathbf{r} \propto (1, \dots, 1)^T$ case response $p_x(t)$ represents the "average" subconverter pulse shape, while $\mathbf{p}_{\mathbf{s}}(t)$ models the mismatch and crosstalk between the subconverters. Error terms $e_x(t)$ and $\mathbf{e}_{\mathbf{s}}(t)$ are the components of the quantization error that lie in the subspace occupied by the input $x(t)$ (along $\boldsymbol{\mu}_1$) and in its orthogonal complement, respectively.

C. Frequency-Domain Analysis

Although the system is structurally best described in the time domain, the basic noise-shaping premise is frequency-domain based, and thus the design of $\Delta\Sigma$ systems is almost always done in the frequency domain. To be somewhat rigorous, this requires that we take one of two approaches. The first is to consider deterministic, finite-energy signals and look at the usual Fourier-transform representations. The other is to model the input, quantizer error, and output as random processes and work with the power spectra of signals of interest. Here we choose the latter, making the simplifying assumptions that $x[n]$, $\mathbf{e}_x[n]$, $\mathbf{e}_{\mathbf{s}}[n]$, and $\boldsymbol{\epsilon}(t)$ are all uncorrelated with each other. Further we assume that $\mathbf{e}_{\mathbf{s}}[n]$ is circular, so that $\mathbf{S}_{e_{\mathbf{s}}}(fT) = S_{e_{\mathbf{s}}}(fT)\mathbf{I}$, and we set $\mathbf{h}_{\mathbf{s}}[n] = h_{\mathbf{s}}[n]\mathbf{I}$ in respect for that circularity. We assume no prior knowledge of $\mathbf{p}_{\mathbf{s}}(t)$ other than perhaps some general sense of its overall magnitude, so for lack of specific vector directions or spectral regions to favor by intelligent design choice, we set $\mathbf{h}_{x\mathbf{s}}[n]$ and $\mathbf{h}_{\mathbf{s}x}[n]$ to zero. Thus the output power spectrum is [13, Preliminaries and Appendix]

$$\begin{aligned}S_y(f) &= |P_x(f)|^2 T S_x(fT) + \mathbf{r}^T\mathbf{S}_{\boldsymbol{\epsilon}}(f)\mathbf{r} \\ &\quad + |H_x(fT)|^2 |P_x(f)|^2 T S_{e_x}(fT) \\ &\quad + |H_{\mathbf{s}}(fT)|^2 \|\mathbf{P}_{\mathbf{s}}(f)\|^2 T S_{e_{\mathbf{s}}}(fT)\end{aligned} \quad (3)$$

where the first term is the desired signal, shaped by the average DAC pulse, the second term is is the additive noise of the subconverters (generally assumed to be DC-only), the third term is the shaped noise introduced by quantizing the signal, and the last term is the shaped mismatch error of the subconverters.

Written in this way, it may appear that the $\Delta\Sigma$ modulator and mismatch-shaping DAC can be considered independent entities, in the spirit of [1,2,3,9]. In general, however, Fig. 1 is not equivalent to a stand-alone modulator feeding a mismatch-shaping DAC, since the error signals $e_x[n]$ and $\mathbf{e}_{\mathbf{s}}[n]$ are related through the quantizer. The potential benefit of combining the modulator and DAC comes from choosing the combined quantizer and customizing the filtering appropriately, which is the topic of the next section.

III. DESIGN ISSUES

The design of a multibit $\Delta\Sigma$ DAC consists of choosing the DAC elements, the quantizer output constellation (a subset of the possible subconverter inputs), a quantization rule for assigning the subconverter inputs, and the error shaping response $\mathbf{h}[n]$. Ideally, we would directly optimize the system parame-

ters, (the quantizer rule and error shaping) to achieve, for example, minimum error energy in the signal passband:

$$\begin{aligned} \underset{\mathbf{Q},\mathbf{h}}{\text{minimize}} \quad & \int_{\mathcal{F}_{\text{passband}}} |\Xi(f)|^2 \, df \\ \text{s.t.} \quad & \text{modulator adequately stable,} \end{aligned}$$

where $\Xi(f)$ is the last two terms of (3). Unfortunately, the highly nonlinear nature of the quantizer makes this an impractical task (such optimal parameters are elusive even for $\Delta\Sigma$ modulators alone), and further $\Xi(f)$ (and thus the optimal system parameters) depends on the input signal through its effect on $\mathbf{e}_x[n]$ and $\mathbf{e}_\mathbf{s}[n]$. Instead, we present a general single-parameter class of vector quantizers, including two special cases of interest, to control the relative magnitudes of errors $e_x[n]$ and $\mathbf{e}_\mathbf{s}[n]$. We then approach the design of the error-shaping filters as an extension of existing $\Delta\Sigma$ literature.

A. Quantizer Design

We define a quantizer by the pair $(\mathcal{C}, \mathbf{Q})$, where constellation $\mathcal{C} = \{\mathbf{c}_k\} \subset \mathbb{R}^M$ is a finite set of possible output vectors and quantizer rule $\mathbf{Q} : \mathbb{R}^M \to \mathcal{C}$ maps inputs to outputs. The inverse images (decision regions) $\mathbf{D}_k = \{\mathbf{d} \in \mathbb{R}^M : \mathbf{Q}(\mathbf{d}) = \mathbf{c}_k\}$ induce a partition $\mathcal{D} = \{\mathbf{D}_k\}$ of $\mathbb{R}^M$. A most-common rule is to map inputs to the nearest constellation point, or equivalently to minimize the Euclidean norm of the error:

$$\mathbf{Q}(\mathbf{d}) = \underset{\mathbf{c}_k \in \mathcal{C}}{\operatorname{argmin}} \|\mathbf{c}_k - \mathbf{d}\|.$$

Here we wish to control the relative magnitudes of the e_x and $\mathbf{e}_\mathbf{s}$ components of the error $\mathbf{e} = \mathbf{Q}(\mathbf{u}) - \mathbf{u}$, and thus we define a new norm

$$\left\| \begin{pmatrix} x \\ \mathbf{s} \end{pmatrix} \right\|_\alpha = \sqrt{\alpha x^2 + (1-\alpha)\|\mathbf{s}\|^2}\,, \quad 0 < \alpha < 1.$$

As α is varied from zero to one, this norm places increasing weight on the x component of the vector, and the corresponding quantization rule increasingly favors reducing e_x at the expense of $\mathbf{e}_\mathbf{s}$.

For example, let us look at the simplest multi-element DAC possible, composed of two one-bit elements with input values $\{-1, 1\}$. Take

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and denote the input to the DAC with the vector $\mathbf{v} = (v_1, v_2)^T$. The corresponding constellation is shown in Fig. 2 (top), in both the (x, s) and (v_1, v_2) coordinate systems. We see that x can take on three distinct values, thus the DAC is indeed multibit. Figure 2 (bottom) shows the partitions corresponding to $\alpha \to 1$, $\alpha = 2/3$, and $\alpha = 1/2$. The partitions for $\alpha = 1/3$ and $\alpha \to 0$ follow in this case by exchanging x and s.

From Fig. 2 we can see two important special cases, which intuitively extend to higher dimensions. As α approaches one, the error in the x direction dominates so strongly that effectively we first choose the subset of the constellation with the closest x value and then choose within that subset to minimize error in the s direction. This is effectively the quantizer used in previous works [1, 2, 3, 9] describing a scalar $\Delta\Sigma$ modulator feeding an independent DEM DAC, where the first quantization step occurs in the $\Delta\Sigma$, and the second occurs in the DAC. When $\alpha = 1/2$ we have a true nearest-neighbor quantizer. Since $\mathbf{A}$ is orthogonal and the constellation is the Cartesian product of orthogonal subconstellations (those of the subconverters) this is easily implemented by first multiplying the input vector by $\mathbf{A}^{-1}$, quantizing the input to each subconverter independently, and then transforming back by $\mathbf{A}$.

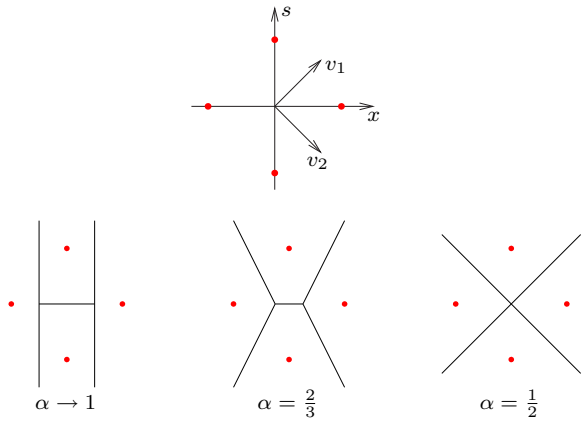

Figure 2: Constellation of a dual unit-element DAC (top), and quantization rules for various values of α.

B. Loop Filtering

Much of the design approach for the loop filtering can be borrowed from standard $\Delta\Sigma$ theory. Based on the quantizer design, approximate quantization error levels can be determined for $e_x[n]$ and $\mathbf{e}_\mathbf{s}[n]$. The individual filters can be determined using cookbook approaches [14] or by optimization [15]. As with standard $\Delta\Sigma$ modulators, the out-of-band gain and the order of the filter are predictors of system stability. Here, however, there are two responses and multiple dimensions, and until $\Delta\Sigma$ theory is extended accordingly, experimentation seems the best way to test for joint stability.

It is not required for $h_x[n]$ and $h_\mathbf{s}[n]$ to have the same order, although empirically this seems to work well. For $\alpha \approx 1$, $h_x[n]$ can be chosen just as for a standalone multibit $\Delta\Sigma$ modulator, with $h_\mathbf{s}[n]$ then tweaked for optimal SNR. Experience has shown [1, 3] that $h_\mathbf{s}[n]$ can generally be chosen as for a one-bit $\Delta\Sigma$ modulator. For $\alpha \approx 1/2$, $h_x[n]$ and $h_\mathbf{s}[n]$ can both be chosen as for a one-bit $\Delta\Sigma$ modulator. Values of α less than $1/2$ make little sense unless the subconverters are multibit DAC's themselves.

IV. DESIGN EXAMPLE

Consider a nine-level $\Delta\Sigma$ DAC comprised of $M = 8$ one-bit subconverters with 1% rms randomly chosen pulse-height mismatch between them, and no timing errors or crosstalk. Let A

be an appropriately scaled 8×8 Hadamard matrix. The constellation $\mathcal{C}$ consists of the 256 corners of an eight-dimensional hypercube. We wish to maximize SNR with an oversampling ratio of 32 and fifth-order noise shaping with optimized zeros. The *Delta-Sigma Toolbox* [14] for Matlab was used to design the noise shaping. We consider $\alpha \to 1$ (the classic solution) and $\alpha = 2/3$.

For $\alpha \to 1$, first $h_x[n]$ was determined by raising the allowed out-of-band gain until the boundary of stability was found with a sinusoidal input. Then the same was done for $h_s[n]$. The resulting subconverter-input vector $\mathbf{v}[n]$ was fed to ideal subconverters, to the mismatched subconverters, and to an unshaped thermometer-encoded DAC (which operates only on the sum of the elements of $\mathbf{v}[n]$ and ignores the mismatch-shaping information). The ideal and thermometer-encoded DACs represent the best- and worst-case performances. Figure 3 shows the output power spectra and SNR for the example (top), as well as the shaped signal-quantization and mismatch-error spectra (bottom). The shaped mismatch error dominates, suggesting a lower value for α.

For $\alpha = 2/3$, the allowable out-of-band gains of the two shaping responses were adjusted to maximize SNR while retaining stability. This was achieved when the two shaped error terms of $\Xi(f)$ were approximately equal[1] in the passband, as seen in Figure 3 (bottom). Reducing α to $2/3$ did in fact lower the mismatch error, but also raised the signal quantization error. Thus, we don't achieve ideal performance, but we do realize another 12 dB of SNR gain over the usual $\Delta\Sigma$ followed by an independent mismatch-shaping DAC. Both mismatch shaping methods are far superior to the unshaped DAC.

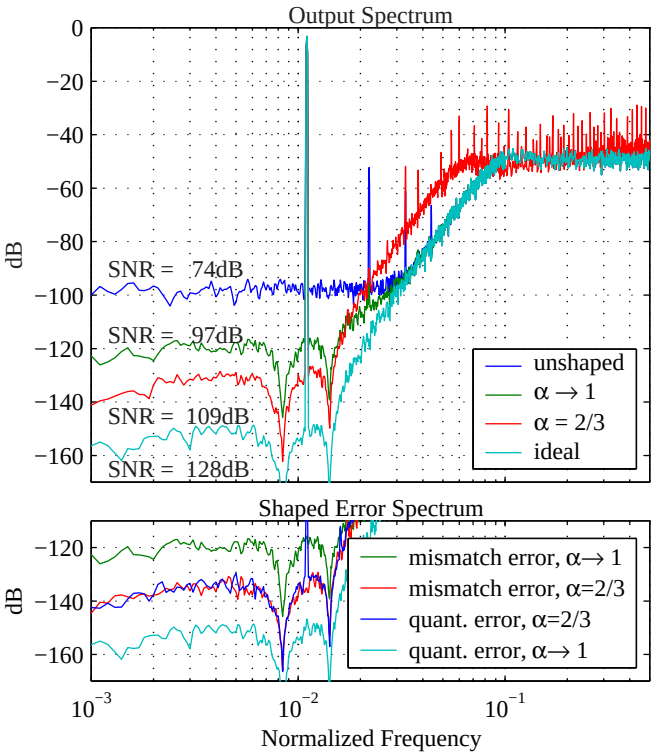

Figure 3: Example output and error spectra.

V. CONCLUSIONS

We have presented a general scalar $\Delta\Sigma$ DAC architecture that allows the spectral shaping of both quantization and hardware-mismatch errors. A simple example demonstrates how adjusting the quantization rule can "even out" the distribution of errors and improve overall SNR. Conventional $\Delta\Sigma$ noise shaping design techniques appear to carry over well to the joint-shaping approach.

REFERENCES

[1] J. O. Coleman and D. P. Scholnik, "Vector switching generalizes D/A noise shaping," in *Midwest Symposium on Circuits and Systems*, (Las Cruces, NM), Aug. 1999.

[2] I. Galton, "Spectral shaping of circuit errors in digital-to-analog convertors," *IEEE Trans. Circuits and Systems II*, vol. 44, pp. 808–817, Oct. 1997.

[3] R. Schreier and B. Zhang, "Noise-shaped multibit D/A convertor employing unit elements," *Electronics Letters*, vol. 31, pp. 1712–1713, Sept. 1995.

[4] R. T. Baird and T. S. Fiez, "Linearity enhancement of multi-bit $\Delta\Sigma$ A/D and D/A converters using data weighted averaging," *IEEE Trans. Circuits and Systems II*, vol. 42, pp. 753–762, Dec. 1995.

[5] L. R. Carley, "A noise-shaping coder topology for 15+ bit converters," *IEEE Journal of Solid-State Circuits*, vol. 24, pp. 267–273, Apr. 1989.

[1]The spectral peak in the shaped quantization error indicates that it is somewhat correlated with the input signal, which can be handled with a cross-correlation term [1] in (3) that effectively modifies $P_x(f)$.

[6] O. J. A. P. Nys and R. K. Henderson, "An analysis of dynamic element matching techniques in sigma-delta modulation," in *Proc. IEEE Int. Symp. Circuits and Systems*, pp. 231–234, 1996.

[7] B. H. Leung and S. Sutaria, "Multibit $\Sigma-\Delta$ A/D convertor incorporating a novel class of dynamic element matching techniques," *IEEE Trans. Circuits and Systems II*, vol. 39, pp. 35–51, Jan. 1992.

[8] D. Cini, C. Samori, and A. L. Lacaita, "Double-index averaging: A novel technique for dynamic element matching in Σ-Δ A/D convertors," *IEEE Trans. Circuits and Systems II*, vol. 46, pp. 353–358, Apr. 1999.

[9] L. Hernández, "A model of mismatch-shaping D/A conversion for linearized DAC architectures," *IEEE Trans. Circuits and Systems I*, vol. 45, pp. 1068–1076, Oct. 1998.

[10] D. P. Scholnik and J. O. Coleman, "Vector delta-sigma modulation with integral shaping of hardware-mismatch errors," in *Proc. IEEE Int. Symp. Circuits and Systems*, (Geneva, Switzerland), May 2000.

[11] R. J. van de Plassche, "Dynamic element matching for high-accuracy monolithic D/A converters," *IEEE Journal of Solid-State Circuits*, vol. 11, pp. 795–800, Dec. 1976.

[12] J. O. Coleman, "Mathematical unification of dynamic-element-matching methods for spectral shaping of hardware-mismatch errors," in *Midwest Symposium on Circuits and Systems*, (Lansing, Mi), Aug. 2000.

[13] J. O. Coleman, "The power spectrum of a generalization of full-response CPM," in *Proc. 2000 Eureopean Signal Processing Conf. (EUSIPCO 2000)*, Sept. 2000.

[14] R. Schreier, *The Delta-Sigma Toolbox 5.2*, Nov. 1999. ftp://ftp.mathworks.com/pub/contrib/v5/control/delsig.

[15] U. Horbach and M. Lang, "Design and implementaiton of sigma-delta D/A converters with optimized loop filters," in *Proc. IEEE Int. Symp. Circuits and Systems*, (New York, NY), Apr. 1991.