

# VECTOR DELTA-SIGMA MODULATION WITH INTEGRAL SHAPING OF HARDWARE-MISMATCH ERRORS

Dan P. Scholnik and Jeffrey O. Coleman

scholnik@nrl.navy.mil  
jeffc@alum.mit.edu

Naval Research Laboratory  
Radar Division  
Washington, DC

## ABSTRACT

A common architecture here unifies ordinary scalar delta-sigma modulators, vector I/O delta-sigma modulators (including parallel banks of scalar ones), mismatch-shaping DACs, delta-sigma modulators incorporating mismatch-shaping DACs in the conventional way, and something new: scalar or vector delta-sigma modulators incorporating spectral shaping of hardware mismatches as a natural and integral part of the modulator structure. The common architecture resembles a conventional delta-sigma loop but using vector signals and a vector quantizer to operate in a space of higher dimension than the input/output, which are inserted/extracted from the loop along a subspace. The expansion of dimension provides the redundancy that permits the spectral shaping of hardware errors. This common view of these systems suggests that the multidimensional structure of the quantizer is key.

## 1. INTRODUCTION

Recently, DAC architectures have been introduced [1–4] that use multiple  $\Delta\Sigma$ -style shaping loops to move hardware-mismatch errors out of the signal band. Such DACs can be used with conventional multibit  $\Delta\Sigma$  modulators to provide the extremely high in-band precision needed. The similarity in the structure of the scalar-input, scalar-output  $\Delta\Sigma$  modulator and the scalar-input, vector-output DAC suggests that a single architecture can be constructed combining the two. A further generalization admits vector inputs and outputs, allowing (for example) conversion of complex signals.

The need for mismatch-shaping DACs arises from the use of multibit  $\Delta\Sigma$  modulation to lower oversampling ratios and quantization noise and to improve stability. These 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), in which

---

This work was partially supported by the AMRFS program (ONR 31) of the Office of Naval Research. The remainder was supported by the Office of Naval Research through its Base program at the Naval Research Laboratory.

$M$  scalar (generally one bit) DAC elements are used to implement a single multibit DAC by summing the outputs. Since the ideal output depends on how many, and not which, elements are enabled, output values can have several equivalent realizations. Switching between these realizations as time progresses has the effect of shaping the mismatch error that arises due to differences between the elements.

A word on notation: for the sake of readability, the explicit dependence on frequency in the functional representation of signals is suppressed in the mathematics that follow. Thus the input signal in Fig. 1 is denoted simply  $\mathbf{x}$ , and not  $\mathbf{x}(f)$ . For system components, frequency dependence is shown in the figure but omitted in the math. A frequency domain representation is used solely to avoid clumsy convolution notation.

## 2. SYSTEM ANALYSIS

### 2.1. Architecture and Hardware Modeling

Figure 1 illustrates the proposed architecture. Conceptually it is similar to a conventional  $\Delta\Sigma$  loop, but with vector signals, a vector quantizer, and a matrix noise transfer function. A key difference is that the quantizer and loop filter are in general of a higher dimension than the input and output, with nonsquare matrix  $\mathbf{R}$  coupling the input and output to the rest of the system. The resulting redundancy is the key to the shaping of hardware errors in the DAC stage. Here,  $\mathbf{x}$  and  $\mathbf{z}$  are  $N \times 1$  vectors,  $\mathbf{y}$  is an  $M \times 1$  vector, and  $\mathbf{A} = (\mathbf{R} \mathbf{S})$  is an  $M \times M$  matrix with orthogonal columns. Matrix  $\mathbf{R}$  has dimension  $M \times N$  and orthonormal columns, so that  $\mathbf{R}^T \mathbf{R} = \mathbf{I}$ . The error-feedback structure shown with  $M \times M$  loop filter response  $\mathbf{A} \mathbf{H} \mathbf{A}^{-1} - \mathbf{I}$  simplifies the analysis, but any of the usual topologies could be used. Note that for this system to be consistent and realizable the loop filter output  $(\mathbf{A} \mathbf{H} \mathbf{A}^{-1} - \mathbf{I})\mathbf{e}$  must depend only on past inputs, or equivalently

$$\mathbf{h}[n] = \mathbf{0}, \quad n < 0 \\ = \mathbf{I}, \quad n = 0$$

where  $\mathbf{h}$  is the inverse Fourier transform of  $\mathbf{H}$ .

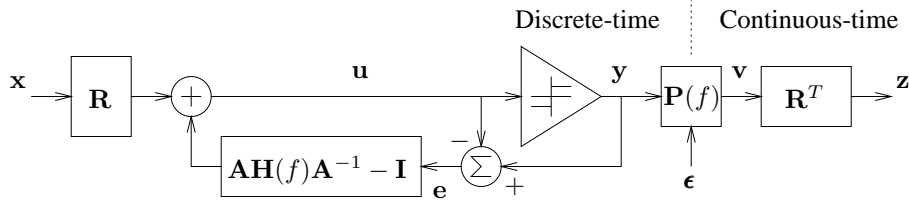


Figure 1: A general vector  $\Delta$ - $\Sigma$  error-feedback architecture incorporating a dynamic element matching (DEM) DAC output.

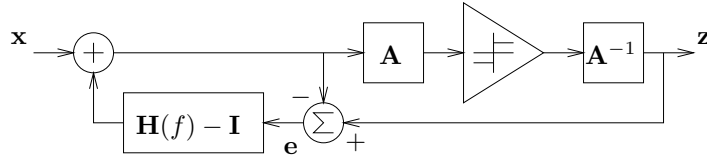


Figure 2: An equivalent system for a vector  $\Delta$  $\Sigma$  modulator without mismatch shaping.

The DAC hardware model used is that of [1], which describes a bank of scalar DAC elements that are combined to form a single multibit DAC. The DAC is defined by the matrix pulse frequency response  $\mathbf{P}$  and the error signal frequency response vector  $\boldsymbol{\epsilon}$ , which in the case of conventional (unshaped) DAC elements is just a constant offset. Element  $\mathbf{P}_{i,j}$  is the response of the  $i$ th element to the  $j$ th input. If all elements have identical output pulses and no crosstalk, then  $\mathbf{P} = P\mathbf{I}$  for a scalar pulse response  $P$ . Any variation along the main diagonal entries is a result of pulse height, pulse shape, and pulse offset differences. Any nonzero entries outside of the main diagonal model crosstalk between elements.

The final DAC output is formed by combining the individual DAC elements, corresponding to a matrix multiply (projection) by  $\mathbf{R}^T$ . In most cases this is performed in analog hardware, and the actual multiplication may be by  $\mathbf{R}^T + \Delta$ , where  $\Delta$  is some error matrix. We note that such error can be incorporated into the DAC model by defining a new matrix pulse response

$$\mathbf{P}' = (\mathbf{I} + \mathbf{R}\Delta)\mathbf{P}$$

with error vector

$$\boldsymbol{\epsilon}' = (\mathbf{I} + \mathbf{R}\Delta)\boldsymbol{\epsilon},$$

and thus we can forego any further reference to  $\Delta$  and consider the output projection to be ideal.

## 2.2. Input-Output Relationship

The analysis of the system in Fig. 1 proceeds as follows. The vector quantizer output can be modeled as the sum of a desired term (the quantizer input) and an error term,  $\mathbf{y} = \mathbf{u} + \mathbf{e}$ . The quantizer input in turn can be written

$$\mathbf{u} = \mathbf{R}\mathbf{x} + (\mathbf{A}\mathbf{H}\mathbf{A}^{-1} - \mathbf{I})\mathbf{e},$$

and substituting this results in

$$\mathbf{y} = \mathbf{R}\mathbf{x} + \mathbf{A}\mathbf{H}\mathbf{A}^{-1}\mathbf{e}.$$

The quantized output is fed to a bank of scalar DAC elements, represented by matrix pulse response  $\mathbf{P}$  and additive error vector  $\boldsymbol{\epsilon}$ . The DAC output vector is

$$\begin{aligned} \mathbf{v} &= \mathbf{P}\mathbf{y} + \boldsymbol{\epsilon} \\ &= \mathbf{P}\mathbf{R}\mathbf{x} + \mathbf{P}\mathbf{A}\mathbf{H}\mathbf{A}^{-1}\mathbf{e} + \boldsymbol{\epsilon}. \end{aligned}$$

The system output is  $\mathbf{z} = \mathbf{R}^T\mathbf{v}$ , or

$$\mathbf{z} = \mathbf{R}^T\mathbf{P}\mathbf{R}\mathbf{x} + \mathbf{R}^T\mathbf{P}\mathbf{A}\mathbf{H}\mathbf{A}^{-1}\mathbf{e} + \mathbf{R}^T\boldsymbol{\epsilon}. \quad (1)$$

We can simplify (1) if we partition  $\mathbf{H}$  as

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_x & \mathbf{H}_{xs} \\ \mathbf{H}_{sx} & \mathbf{H}_s \end{pmatrix}.$$

The prior restrictions on the structure of the matrix  $\mathbf{A}$  allow us to write

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} \mathbf{R}^T\mathbf{R} & \mathbf{R}^T\mathbf{S} \\ \mathbf{S}^T\mathbf{R} & \mathbf{S}^T\mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_S \end{pmatrix},$$

with  $\mathbf{D}_S$  a diagonal matrix, and therefore,

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{R}^T \\ \mathbf{D}_S^{-1}\mathbf{S}^T \end{pmatrix}.$$

Equation 1 now becomes

$$\mathbf{z} = \mathbf{P}_x\mathbf{x} + (\mathbf{P}_x \ \mathbf{P}_s) \begin{pmatrix} \mathbf{H}_x & \mathbf{H}_{xs} \\ \mathbf{H}_{sx} & \mathbf{H}_s \end{pmatrix} \begin{pmatrix} \mathbf{e}_x \\ \mathbf{e}_s \end{pmatrix} + \mathbf{R}^T\boldsymbol{\epsilon}, \quad (2)$$

where

$$\begin{aligned} \mathbf{P}_x &\triangleq \mathbf{R}^T\mathbf{P}\mathbf{R} \\ \mathbf{P}_s &\triangleq \mathbf{R}^T\mathbf{P}\mathbf{S} \\ \mathbf{e}_x &\triangleq \mathbf{R}^T\boldsymbol{\epsilon} \\ \mathbf{e}_s &\triangleq \mathbf{D}_S^{-1}\mathbf{S}^T\boldsymbol{\epsilon}. \end{aligned}$$

The matrix response  $\mathbf{P}_x$  represents the “average” DAC pulse shape, while  $\mathbf{P}_s$  models the mismatch and crosstalk between the DAC elements. When  $N > 1$  we can see that  $\mathbf{P}_x$  may well not be ideal, in that it may have variance along the main diagonal and nonzero entries off the diagonal. The resulting gain error and crosstalk is inevitable with multiple analog output channels, and would be present even if each channel was processed separately. Error terms  $e_x$  and  $e_s$  correspond to the components of the quantization error that lie in the subspace occupied by the input  $\mathbf{x}$  (the column space of  $\mathbf{R}$ ) and its orthogonal complement, respectively.

### 2.3. Analysis

The output (2) contains one desired signal term and two error terms. The first term is the desired signal shaped by the nominal pulse-response matrix  $\mathbf{P}_x$ . The third term is just a constant in most cases and will be ignored. The middle term, when multiplied through, results in four error terms. The term  $\mathbf{P}_x \mathbf{H}_x e_x$  represents the quantization error of the input signal, shaped by the  $\mathbf{H}_x$  block of the noise transfer function and the nominal output pulse. This error is analogous to the quantization error in a conventional  $\Delta\Sigma$  modulator and is largely unaffected by hardware mismatch. The  $\mathbf{P}_s \mathbf{H}_s e_s$  term represents the pure shaped mismatch error, which is largely signal independent. The remaining two cross terms can be canceled simply by choosing  $\mathbf{H}_{sx} = \mathbf{H}_{xs}^T = \mathbf{0}$ . If  $\mathbf{P}$  were known (in general it is not) then these terms might be chosen differently to cancel part of the quantization and mismatch terms.

The key to the design is the choice of noise transfer functions  $\mathbf{H}_x$  and  $\mathbf{H}_s$  and the quantization rule. As mentioned previously, response  $\mathbf{H}_x$  shapes the signal quantization error, which is a reasonably well known quantity that depends mainly on the quantizer characteristic. The mismatch error, on the other hand, is controlled by response  $\mathbf{H}_s$ , and that error magnitude depends on both the quantizer and the total amount of hardware mismatch as characterized by  $\mathbf{P}_s$ . As in conventional  $\Delta\Sigma$  theory, how aggressive the shaping can be while maintaining stability is controlled by the quantization. Because the quantizer operates in a higher dimension than that of the input signal, there is no unique way to choose a decision rule, and we have some control on which subspaces contain the greatest errors. When hardware errors are small, we might first nearest-neighbor quantize just the signal subspace in order to minimize signal quantization error  $e_x$  (at the expense of greater error  $e_s$ ). When hardware errors are large, we might choose a true nearest-neighbor quantizer, or even try to reduce  $e_s$  by increasing  $e_x$ .

## 3. SPECIAL CASES

### 3.1. Vector $\Delta\Sigma$ Modulators

For  $\Delta\Sigma$  modulators with digital outputs (and thus no error shaping), we have  $\mathbf{P} = \mathbf{I}$ ,  $\epsilon = 0$ , and  $\mathbf{A} = \mathbf{R} = \mathbf{A}^{-T}$ , which simplifies the output expression considerably:

$$\begin{aligned} \mathbf{z} &= \mathbf{x} + \mathbf{H}\mathbf{A}^{-1}\mathbf{e} \\ &= \mathbf{x} + \mathbf{H}_x e_x. \end{aligned}$$

As in the usual scalar  $\Delta\Sigma$  model, the output is the sum of the input and a noise term shaped by a given transfer function. In general the input and output are vectors, and the error-shaping transfer function is a matrix. In most cases we don’t care about the directional response of  $\mathbf{H}$  and thus set  $\mathbf{H} = H\mathbf{I}$ , where  $H$  is a scalar transfer function.

Figure 2 depicts an equivalent system for a vector  $\Delta\Sigma$  modulator without mismatch shaping, which shows that matrix  $\mathbf{A}$  affects only the operation of the quantizer and can in fact be absorbed into it without loss of generality. If the resulting quantizer decides each element’s input independently and  $\mathbf{H}$  is diagonal, the result is just  $N$  independent scalar  $\Delta\Sigma$  modulators operating in parallel. If the quantizer output constellation is a Cartesian product of  $N$  identical one-dimensional constellations and  $\mathbf{H} = H\mathbf{I}$  the scalar modulators are all identical.

### 3.2. Scalar Error-Shaping DAC

References [1–4] all describe mismatch-shaping DACs that use the output of a multibit scalar  $\Delta\Sigma$  as the input, and the sum of  $M$  elements as the output. This becomes here the special case of  $N = 1$  and  $\mathbf{R} = [1 \ 1 \ \dots \ 1]^T / \sqrt{M}$ . The rest of the matrix  $\mathbf{A}$  can be chosen somewhat arbitrarily to satisfy the column-orthogonality requirements, with the obvious choice being an orthogonal  $\mathbf{A}$ . In at least one case, however, the use of a nonorthogonal  $\mathbf{A}$  (the columns of  $\mathbf{S}$  are orthogonal but not orthonormal) leads to (or, depending on the viewpoint, stems from) a particularly simple and efficient hardware implementation [2].

Let us consider the most-common case where the  $M$  elements each have one bit of resolution. With  $\mathbf{R}$  as defined above the output has  $M + 1$  possible levels, and we can choose scalar response  $\mathbf{H}_x$  (typically of a high order) based on standard multibit  $\Delta\Sigma$  design techniques. For the mismatch-shaping response assuming no prior knowledge of  $\mathbf{P}_s$  we can choose  $\mathbf{H}_s = H_s \mathbf{I}$ , with  $H_s$  a (lower-order) response that would produce a stable one-bit  $\Delta\Sigma$  modulator. Since the shaping of signal quantization errors is more aggressive than the shaping of mismatch errors, we choose a quantizer that first decides along the  $\mathbf{R}$  direction (selecting the nearest output level) and then implements a nearest-neighbor rule on the resulting smaller constellation. This two-step decision ensures the minimum signal quantization

noise level at the expense of increased mismatch error, which is appropriate when the quantization error dominates but may be quite suboptimal if hardware errors are large. Note that when the signal and mismatch components are independently filtered and quantized in this manner the two processes can be completely separated, resulting in the cascade of a scalar multibit  $\Delta\Sigma$  modulator and mismatch-shaping DAC of [1–3]. Further, if the input signal is already properly quantized the fed-back error will be zero, and the system will act solely as a mismatch-shaping DAC.

#### 4. CONCLUSIONS

In this paper we describe a new generalized vector architecture to encompass and extend current  $\Delta\Sigma$  systems incorporating hardware mismatch-shaping DACs. There are many open design issues left to explore for the fully general vector modulator. It seems probable that, given some knowledge of the magnitude of the hardware mismatch errors, the quantization rule can be chosen to optimally distribute the quantization error across the signal and mismatch subspaces. While the design of a noise transfer function for a scalar  $\Delta\Sigma$  modulator is well covered in the literature, it is not obvious how to extend these approaches to the combined shaping  $\mathbf{H}$  of the general system to minimize the total in-band quantization noise. Such an extension might offer a significant improvement over the (often unsatisfactory) ad hoc choosing of  $\mathbf{H}_x$  and  $\mathbf{H}_x$  independently. Also open is how the choice of quantizer affects the choice of  $\mathbf{H}$ .

#### 5. REFERENCES

- [1] J. O. Coleman and D. P. Scholnik, “Vector switching generalizes D/A noise shaping,” in *Proc. Midwest Symp. on Circuits and Systems (MWSCAS)*, 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, no. 10, pp. 808–817, Oct. 1997.
- [3] R. Schreier and B. Zhang, “Noise-shaped multibit D/A convertor employing unit elements,” *Electronics Letters*, vol. 31, no. 20, 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.