# IIR COEFFICIENT QUANTIZATION ANALYSIS<sup>\*</sup>

Douglas L. Jones

This work is produced by OpenStax-CNX and licensed under the Creative Commons Attribution License  $1.0^{\dagger}$ 

#### Abstract

Proper coefficient quantization is essential for IIR filters. Sensitivity analysis shows that second-order cascade-form implementations have much lower sensitivity than higher-order direct-form or transpose-form structures. The normal form is even less sensitive, but requires more computation.

Coefficient quantization is an important concern with IIR filters, since straigthforward quantization often yields poor results, and because quantization can produce unstable filters.

## 1 Sensitivity analysis

The performance and stability of an IIR filter depends on the pole locations, so it is important to know how quantization of the filter coefficients  $a_k$  affects the pole locations  $p_i$ . The denominator polynomial is

$$D(z) = 1 + \sum_{k=1}^{N} a_k z^{-k} = \prod_{i=1}^{N} 1 - p_i z^{-1}$$

We wish to know  $\frac{\partial p_i}{\partial a_k}$ , which, for small deviations, will tell us that a  $\delta$  change in  $a_k$  yields an  $\epsilon = \delta \frac{\partial p_i}{\partial a_k}$ change in the pole location.  $\frac{\partial p_i}{\partial a_k}$  is the **sensitivity** of the pole location to quantization of  $a_k$ . We can find  $\frac{\partial p_i}{\partial a_k}$  using the chain rule.

$$\frac{\partial A(z)}{\partial a_k}|_{z=p_i} = \frac{\partial A(z)}{\partial z} \frac{\partial z}{\partial a_k}|_{z=p}$$

$$\frac{\partial p_i}{\partial a_k} = \frac{\frac{\partial A(z_i)}{\partial a_k}|_{z=p_i}}{\frac{\partial A(z_i)}{\partial z}|_{z=p_i}}$$

which is

$$\frac{\partial p_i}{\partial a_k} = \frac{z^{-k}}{-(z^{-1}\prod_{j=j\neq i,1}^N 1 - p_j z^{-1})}|_{z=p_i}$$

$$= \frac{-p_i^{N-k}}{\prod_{j=j\neq i,1}^N p_j - p_i}$$
(1)

\*Version 1.2: Jan 2, 2005 2:25 pm -0600

<sup>†</sup>http://creativecommons.org/licenses/by/1.0

Note that as the poles get closer together, the sensitivity increases greatly. So as the filter order increases and more poles get stuffed closer together inside the unit circle, the error introduced by coefficient quantization in the pole locations grows rapidly.

How can we reduce this high sensitivity to IIR filter coefficient quantization?

#### 1.1 Solution

Cascade<sup>1</sup> or parallel form<sup>2</sup> implementations! The numerator and denominator polynomials can be factored off-line at very high precision and grouped into second-order sections, which are then quantized section by section. The sensitivity of the quantization is thus that of second-order, rather than N-th order, polynomials. This yields major improvements in the frequency response of the overall filter, and is almost always done in practice.

Note that the numerator polynomial faces the same sensitivity issues; the **cascade** form also improves the sensitivity of the zeros, because they are also factored into second-order terms. However, in the **parallel** form, the zeros are globally distributed across the sections, so they suffer from quantization of all the blocks. Thus the **cascade** form preserves zero locations much better than the parallel form, which typically means that the stopband behavior is better in the cascade form, so it is most often used in practice.

NOTE: On the basis of the preceding analysis, it would seem important to use cascade structures in FIR filter implementations. However, most FIR filters are linear-phase and thus symmetric or anti-symmetric. As long as the quantization is implemented such that the filter coefficients retain symmetry, the filter retains linear phase. Furthermore, since all zeros off the unit circle must appear in groups of four for symmetric linear-phase filters, zero pairs can leave the unit circle only by joining with another pair. This requires relatively severe quantizations (enough to completely remove or change the sign of a ripple in the amplitude response). This "reluctance" of pole pairs to leave the unit circle tends to keep quantization from damaging the frequency response as much as might be expected, enough so that cascade structures are rarely used for FIR filters.

#### Exercise 1

What is the worst-case pole pair in an IIR digital filter?

## 2 Quantized Pole Locations

In a direct-form<sup>3</sup> or transpose-form<sup>4</sup> implementation of a second-order section, the filter coefficients are quantized versions of the polynomial coefficients.

$$D(z) = z^2 + a_1 z + a_2 = (z - p) (z - \overline{p})$$
$$p = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}$$
$$p = re^{i\theta}$$

$$D(z) = z^2 - 2r\cos(\theta) + r^2$$

(Solution on p. 6.)

<sup>&</sup>lt;sup>1</sup>"IIR Filter Structures": Section IIR Cascade Form <a href="http://cnx.org/content/m11919/latest/#section6">http://cnx.org/content/m11919/latest/#section6</a>>

 $<sup>\</sup>label{eq:linear} \ensuremath{^2"\text{IIR Filter Structures": Section Parallel form < http://cnx.org/content/m11919/latest/\#section20>} \ensuremath{^2"\text{IIR Filter Structures": Section Parallel form < http://cnx.org/content/m11919/latest/#section20>} \ensuremath{^2"\text{IIR Filter Structures: Section Parallel form < http://cnx.org/content/m11919/latest/#section20>} \ensuremath{^2"\text{IIR Filter Structures: Section Parallel form < http://cnx.org/content/m11919/latest/#section20>} \ensuremath{^2"\text{IIR Filter Structures: Section Parallel form < http://section20>} \ensuremath{^2"\text{IIR Filter Structures: Section Parallel form < http://section20$ 

<sup>&</sup>lt;sup>3</sup>"IIR Filter Structures": Section Direct-form I IIR Filter Structure <a href="http://cnx.org/content/m11919/latest/#section1">http://cnx.org/content/m11919/latest/#section1</a> <sup>4</sup>"IIR Filter Structures": Section Transpose-Form IIR Filter Structure

The Filter Structures": Section Transpose-Form The Filter Struct

 $<sup>&</sup>lt;\!http://cnx.org/content/m11919/latest/\#section4\!>$ 

 $\operatorname{So}$ 

$$a_1 = -\left(2r\cos\left(\theta\right)\right)$$
$$a_2 = r^2$$

Thus the quantization of  $a_1$  and  $a_2$  to B bits restricts the radius r to  $r = \sqrt{k\Delta_B}$ , and  $a_1 = -(2\Re(p)) = k\Delta_B$ . The following figure shows all stable pole locations after four-bit two's-complement quantization.



Figure 1

Note the nonuniform distribution of possible pole locations. This might be **good** for poles near r = 1,  $\theta = \frac{\pi}{2}$ , but not so good for poles near the origin or the Nyquist frequency.

In the "normal-form" structures, a state-variable<sup>5</sup> based realization, the poles are uniformly spaced.

 $<sup>^{-5}</sup>$ "State-Variable Representation of Discrete-Time Systems", Definition 1: "State"  $< \rm http://cnx.org/content/m11920/latest/\#state>$ 



Figure 2

This can only be accomplished if the coefficients to be quantized equal the real and imaginary parts of the pole location; that is, 10

$$\alpha_1 = r\cos\left(\theta\right) = \Re\left(r\right)$$

$$\alpha_2 = r\sin\left(\theta\right) = \Im\left(p\right)$$

This is the case for a 2nd-order system with the state matrix  ${}^{6}A = \begin{pmatrix} \alpha_{1} & \alpha_{2} \\ -\alpha_{1} & \alpha_{1} \end{pmatrix}$ : The denominator

polynomial is

$$det (zI - A) = (z - \alpha_1)^2 + \alpha_2^2$$
  
=  $z^2 - 2\alpha_1 z + \alpha_1^2 + \alpha_2^2$   
=  $z^2 - 2r\cos(\theta) z + r^2 (\cos^2(\theta) + \sin^2(\theta))$   
=  $z^2 - 2r\cos(\theta) z + r^2$  (2)

<sup>&</sup>lt;sup>6</sup>"State-Variable Representation of Discrete-Time Systems": Section State and the State-Variable Representation <http://cnx.org/content/m11920/latest/#section1>

Given any second-order filter coefficient set, we can write it as a state-space system<sup>7</sup>, find a transformation matrix<sup>8</sup>T such that  $A = T^{-1}AT$  is in normal form, and then implement the second-order section using a structure corresponding to the state equations.

The normal form has a number of other advantages; both eigenvalues are equal, so it minimizes the norm of Ax, which makes overflow less likely, and it minimizes the output variance due to quantization of the state values. It is sometimes used when minimization of finite-precision effects is critical.

### Exercise 2

What is the disadvantage of the normal form?

(Solution on p. 6.)

 $<sup>^7</sup>$  "State-Variable Representation of Discrete-Time Systems": Section State and the State-Variable Representation  $<\!http://cnx.org/content/m11920/latest/#section1>$ 

 $<sup>^{8}</sup>$  "State-Variable Representation of Discrete-Time Systems": Section State-Variable Transformation  $<\!http://cnx.org/content/m11920/latest/\#statetrans>$ 

## Solutions to Exercises in this Module

## Solution to Exercise (p. 2)

The pole pair closest to the real axis in the z-plane, since the complex-conjugate poles will be closest together and thus have the highest sensitivity to quantization.

#### Solution to Exercise (p. 5)

It requires more computation. The general state-variable equation<sup>9</sup> requires nine multiplies, rather than the five used by the Direct-Form  $II^{10}$  or Transpose-Form<sup>11</sup> structures.

<sup>&</sup>lt;sup>9</sup>"State-Variable Representation of Discrete-Time Systems", Definition 1: "State" <http://cnx.org/content/m11920/latest/#state>

<sup>&</sup>lt;sup>10</sup>"IIR Filter Structures": Section Direct-Form II IIR Filter Structure <a href="http://cnx.org/content/m11919/latest/#section3">http://cnx.org/content/m11919/latest/#section3</a> <sup>11</sup>"IIR Filter Structures": Section Transpose-Form IIR Filter Structure

 $<sup>&</sup>lt;\! http://cnx.org/content/m11919/latest/\#section4\!>$