Abstract—This paper investigates the problem of reducing the deviation from a desired transfer function caused by the coefficient quantization errors of a three-dimensional (3-D) separable-denominator digital filter. First, a 3-D transfer function with separable denominator is represented with the cascade connection of three one-dimensional (1-D) transfer functions by applying a minimal decomposition technique. Next, the multi-input multi-output (MIMO) 1-D transfer function located in the middle of the cascade connection is realized by a minimal state-space model and then the $l_2/l_2$-sensitivity of the model is analyzed. Third, a technique for the optimal synthesis of the minimal state-space model is developed so as to minimize the $l_2$-sensitivity subject to $l_2$-scaling constraints. Finally, a numerical example is given to demonstrate the validity and effectiveness of the proposed technique.

I. INTRODUCTION

The problem of minimizing the coefficient sensitivity for two-dimensional (2-D) state-space digital filters has been explored extensively [1]-[7]. Several techniques have been proposed for synthesizing 2-D state-space filter structures with minimum coefficient sensitivity [2]-[7]. Some of them evaluate the sensitivity by using a mixture of $l_1/l_2$ norms [1]-[4]. The others rely on the use of a pure $l_2$ norm [5]-[7]. In [6], the weighted-sensitivity minimization of 2-D state-space digital filters has been considered in both cases of a mixture of $L_1/L_2$ norms and a pure $L_2$ norm. It should be noted that the $l_2$-sensitivity minimization is more natural and reasonable than the conventional $l_1/l_2$ mixed sensitivity minimization, but it is more challenging [7]. A technique has also been proposed for synthesizing three-dimensional (3-D) separable-denominator (SD) state-space digital filters with minimum $l_2$-sensitivity [8]. More recently, the minimization problem of $l_2$-sensitivity subject to $l_2$-scaling constraints has been treated for 2-D state-space digital filters [9],[10]. It is known that the use of scaling constraints can be beneficial for suppressing overflow [11]. Alternatively, 3-D digital filters find applications in various image and video signal processing problems, and the coefficient sensitivity of these filters is directly related to their performance on finite word-length devices.

This paper treats the realization of 3-D SD digital filters which reduce $l_2$-sensitivity subject to $l_2$-scaling constraints. First, a 3-D transfer function with separable denominator is decomposed into three one-dimensional (1-D) transfer functions with a cascade connection, and the MIMO 1-D transfer function in the middle of the cascade connection is described by minimal state-space realization. Next, an iterative procedure is developed for constructing the optimal state-space model for the MIMO 1-D system so as to minimize the $l_2$-sensitivity subject to $l_2$-scaling constraints. Finally, a numerical example is given to demonstrate the validity and effectiveness of the proposed technique.

II. PROBLEM STATEMENT

Consider a stable 3-D SD digital filter described by

$$H(z_1, z_2, z_3) = \frac{N(z_1, z_2, z_3)}{D_1(z_1)D_2(z_2)D_3(z_3)}$$  (1)

where

$$N(z_1, z_2, z_3) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \sum_{k=0}^{N_3} a_{ijk} z_1^{-i} z_2^{-j} z_3^{-k}$$

$$D_1(z_1) = 1 + b_{11} z_1^{-1} + \cdots + b_{1N_1} z_1^{-N_1}$$

$$D_2(z_2) = 1 + b_{21} z_2^{-1} + \cdots + b_{2N_2} z_2^{-N_2}$$

$$D_3(z_3) = 1 + b_{31} z_3^{-1} + \cdots + b_{3N_3} z_3^{-N_3}.$$  

The 3-D transfer function in (1) can be decomposed into three 1-D systems as

$$H(z_1, z_2, z_3) = \frac{Z_1^T}{D_1(z_1)} H_2(z_2) \frac{Z_3}{D_3(z_3)}$$  (2)

where

$$Z_1 = (1, z_1^{-1}, \cdots, z_1^{-N_1})^T$$

$$Z_3 = (1, z_3^{-1}, \cdots, z_3^{-N_3})^T$$

$$H_2(z_2) = \frac{\Delta_0 + \Delta_1 z_2^{-1} + \cdots + \Delta_N z_2^{-N}}{D_2(z_2)}$$

$$\Delta_m = \begin{bmatrix} a_{0m0} & a_{0m1} & \cdots & a_{0mN_3} \\ a_{1m0} & a_{1m1} & \cdots & a_{1mN_3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{Nm0} & a_{Nm1} & \cdots & a_{NmN_3} \end{bmatrix}$$

$m = 0, 1, \cdots, N_2$.

The above 1-D transfer function $H_2(z_2)$ with $(N_3 + 1)$ inputs and $(N_1 + 1)$ outputs can be realized by the following minimal
state-space model \((A_2, B_2, C_2, \Delta_0)p\):
\[
x(k + 1) = A_2 x(k) + B_2 u(k)
y(k) = C_2 x(k) + \Delta_0 u(k)
\] (3)
where \(x(k)\) is a \(p \times 1\) state-variable vector, \(u(k)\) is an \((N_3 + 1) \times 1\) input vector, \(y(k)\) is an \((N_1 + 1) \times 1\) output vector, and \(A_2, B_2, C_2\) and \(\Delta_0\) are real constant matrices of appropriate dimensions. The transfer function of the linear system in (3) is given by
\[
H_2(z) = C_2 (z I_p - A_2)^{-1} B_2 + \Delta_0.
\] (4)

Definition 1: Let \(X\) be an \(m \times n\) real matrix and let \(f(X)\) be a scalar complex function of \(X\), differentiable with respect to all the entries of \(X\). The sensitivity function of \(f(X)\) with respect to \(X\) is then defined as
\[
S_X = \frac{\partial f}{\partial X}, \quad (S_X)_{ij} = \frac{\partial f}{\partial x_{ij}}
\] (5)
where \(x_{ij}\) denotes the \((i, j)\)th entry of matrix \(X\).

Definition 2: Let \(X(z_1, z_2, z_3)\) be an \(m \times n\) complex matrix-valued function of complex variables \(z_1, z_2,\) and \(z_3\). Let \(x_{pq}(z_1, z_2, z_3)\) be the \((p, q)\)th entry of \(X(z_1, z_2, z_3)\). Then the \(l_2\)-norm of \(X(z_1, z_2, z_3)\) is defined as
\[
\|X(z_1, z_2, z_3)\|_2 = \left( \text{tr} \left[ \frac{1}{(2\pi)^d} \int_{|z_1|=1} \int_{|z_2|=1} \int_{|z_3|=1} X(z_1, z_2, z_3) \cdot X^*(z_1, z_2, z_3) \right] \right)^{\frac{1}{2}}.
\] (6)

From (2), (4) and Definitions 1 and 2, the \(l_2\)-sensitivity of the transfer function \(H(z_1, z_2, z_3)\) with respect to the coefficient matrices \(A_2, B_2,\) and \(C_2\) is evaluated by
\[
S = \left\| \frac{\partial H(z_1, z_2, z_3)}{\partial A_2} \right\|_2^2 + \left\| \frac{\partial H(z_1, z_2, z_3)}{\partial B_2} \right\|_2^2 + \left\| \frac{\partial H(z_1, z_2, z_3)}{\partial C_2} \right\|_2^2
\]
\[
= \left\| f(z_2, z_3) g(z_1, z_2) \right\|_2^2 + \left\| \frac{Z_3}{D_3(z_1)} g(z_1, z_2) \right\|_2^2
\]
\[
+ \left\| \frac{f(z_2, z_3)}{D_1(z_1)} \right\|_2^2
\] (7)
where
\[
f(z_2, z_3) = (z_2 I_p - A_2)^{-1} B_2 \frac{Z_3}{D_3(z_3)}
g(z_1, z_2) = \frac{Z_3^T}{D_1(z_1)} C_2 (z_2 I_p - A_2)^{-1}
\]
\[
\frac{Z_3}{D_3(z_3)} = c_1 (z_1 I_{N_1} - A_1)^{-1} B_1 + d_1
\]
\[
\frac{Z_3}{D_3(z_3)} = c_3 (z_3 I_{N_3} - A_3)^{-1} b_3 + d_3.
\]

The \(l_2\)-sensitivity measure in (7) can be written as
\[
S = \text{tr}[M_A(I_p)] + \text{tr}[W_B] + \text{tr}[K_C]
\] (8)
where \(M_A(I_p), W_B,\) and \(K_C\) are obtained by the following general expression:
\[
X = \frac{1}{(2\pi)^d} \int_{|z_1|=1} \int_{|z_2|=1} \int_{|z_3|=1} Y(z_1, z_2, z_3) \cdot Y^*(z_1, z_2, z_3)
\]
\[
\cdot \left( \frac{Z_3}{D_3(z_3)} \right)^* \text{ for } X = M_A(I_p)
\]
\[
Y(z_1, z_2, z_3) = \left[ \frac{Z_3}{D_3(z_3)} g(z_1, z_2) \right]^* \text{ for } X = W_B
\]
\[
Y(z_1, z_2, z_3) = f(z_2, z_3) \frac{Z_3^T}{D_1(z_1)} \text{ for } X = K_C.
\]
The Gramians \(M_A(P), W_B,\) and \(K_C\) can be computed using
\[
M_A(P) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left[ 0 \right]_{ij} \left[ A_2^T \right]_{ij} \frac{A_2^T}{C_2^T R_{ij}^T B_2^T} \frac{0}{A_2^T}
\]
\[
W_B = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left[ A_2^T \right]_{ij} \frac{C_2^T}{D_3(z_3)} R_{ij}^T B_2^T \left( A_2^T \right)_{uj}
\]
\[
K_C = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left[ A_2^T \right]_{ij} \frac{B_2 R_{ij}^T B_2^T}{D_3(z_3)} \left( A_2^T \right)_{uj}
\] (9)
where
\[
\frac{Z_3^T}{D_3(z_1)} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} R_{ij} z_1^{-1} z_3^j
\]
\[
\frac{Z_3^T}{D_3(z_1)} = \sum_{i=0}^{\infty} c_i A_1^i z_1^{-i} + d_1
\]
\[
\frac{Z_3^T}{D_3(z_3)} = \sum_{j=0}^{\infty} C_3 A_3^j z_3^{-j} + d_3.
\]

Applying a coordinate transformation defined by \(\tau(k) = T^{-1} x(k)\) to the 1-D system \((A_2, B_2, C_2, \Delta_0)p\) in (3), we obtain a new realization \((\bar{A}_2, \bar{B}_2, \bar{C}_2, \bar{\Delta}_0)p\) characterized by
\[
\bar{A}_2 = T^{-1} A_2 T, \quad \bar{B}_2 = T^{-1} B_2, \quad \bar{C}_2 = C_2 T.
\] (10)
For the new realization, the \(l_2\)-sensitivity measure in (8) is changed to
\[
S(P) = \text{tr}[M_A(P)P] + \text{tr}[W_B P] + \text{tr}[K_C P^{-1}]
\] (11)
where \(P = TT^T\). Noting that \(f(z_2, z_3)\) is the transfer function from the filter input to the state-variable vector \(x(k)\),
a controllability Gramian $K$ can be derived from

$$
K = \frac{1}{(2\pi)^2} \int_{|z_2|=1} \oint_{|z_3|=1} f(z_2, z_3) f^*(z_2, z_3) \frac{dz_2}{z_2} \frac{dz_3}{z_3}
$$

$$
= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} A_j^T B_j r_j^T B_j^T (A_j^T)^k
$$

where

$$
\frac{Z_3}{D_3(z_3)} = \sum_{j=0}^{\infty} r_j z_3^{-j}.
$$

In this case, $l_2$-scaling constraints are given by

$$(\bar{K})_{ij} = (T^{-1} K T^{-T})_{ij} = 1 \quad \text{for} \quad i = 1, 2, \cdots, p. \quad (13)$$

The problem at hand can now be formulated as to obtain the coordinate transformation matrix $T$ that minimizes $S(P)$ in (11) subject to the $l_2$-scaling constraints in (13).

### III. $L_2$-Sensitivity Minimization

Let the $l_2$-scaling constraints in (13) be relaxed as

$$
\text{tr}[T^{-1} K T^{-T}] = \text{tr}[KP^{-1}] = p. \quad (14)
$$

If $\text{tr}[KP^{-1}] = p$ is satisfied, then a $p \times p$ orthogonal matrix $U$ can always be constructed so that $T = P^{1/2} U$ satisfies $l_2$-scaling constraints in (13) [9]. This justifies the relaxation made in (14) and in this way, we now focus on the problem

$$
\begin{align*}
\text{minimize} & \quad S(P) \quad \text{in} \quad (11) \\
\text{subject to} & \quad \text{tr}[KP^{-1}] = p. \quad (15)
\end{align*}
$$

To solve problem (15), we define the following Lagrange function of the problem:

$$
J(P, \lambda) = \text{tr}[M_A(P) P] + \text{tr}[W_B P]
+ \text{tr}[K C P^{-1}] + \lambda(\text{tr}[KP^{-1}] - p) \quad (16)
$$

where $\lambda$ is the Lagrange multiplier. Setting $\partial J(P, \lambda)/\partial P = 0$, it follows that

$$
P \dot{F}(P) P = G(P, \lambda) \quad (17)
$$

where

$$
\begin{align*}
F(P) &= M_A(P) + W_B \\
G(P, \lambda) &= N_A(P) + K C + \lambda K
\end{align*}
$$

with

$$
N_A(P) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} [I_p \quad 0] \begin{bmatrix} A_2 & B_2 R_3 C_2 \\ 0 & A_2 \end{bmatrix}^k
\begin{bmatrix} 0 & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} A_2^T & 0 \\ C_2^T R_3^T B_2^T & A_2^T \end{bmatrix}^k [I_p \quad 0] .
$$

The equation in (17) is highly nonlinear with respect to $P$. An effective approach to solving the equation in (17) is to relax it into the following recursive second-order matrix equation:

$$
P^{(k+1)} F(P^{(k)}) P^{(k+1)} = G(P^{(k)}, \lambda^{(k+1)}) \quad (18)
$$

where $P^{(k)}$ is assumed to be known from the previous recursion and the solution $P^{(k+1)}$ is given by [9]

$$
P^{(k+1)} = F(P^{(k)})^{-1/2} [F(P^{(k)})^{1/2} G(P^{(k)}, \lambda^{(k+1)}) F(P^{(k)})^{1/2}]^{1/2} F(P^{(k)})^{-1/2}. \quad (19)
$$

Here, the Lagrange multiplier $\lambda^{(k+1)}$ can be efficiently obtained using a bisection method so that

$$
f(\lambda^{(k+1)}) = p - \text{tr}[\dot{K}(k) G(k) (\lambda^{(k+1)})] = 0 \quad (20)
$$

where

$$
\dot{K}^{(k)} = F(P^{(k)}) K F(P^{(k)})^{1/2}
G^{(k)}(\lambda^{(k+1)}) = [F(P^{(k)})^{1/2} G(P^{(k)}, \lambda^{(k+1)}) F(P^{(k)})^{1/2}]^{-1/2}.
$$

This iteration process continues until

$$
|J(P^{(k)}, \lambda^{(k+1)}) - J(P^{(k-1)}, \lambda^{(k)})| < \varepsilon \quad (21)
$$

is satisfied for a prescribed tolerance $\varepsilon > 0$. If the iteration is terminated at step $k$, then $P^{(k)}$ is claimed to be a solution point.

### IV. Numerical Example

Consider a stable 3-D SD digital filter specified by

$$
\Delta_0 = 10^{-2} \begin{bmatrix}
0.00730 & 0.34297 & -0.09594 & 0.20541 \\
3.33408 & -5.73707 & 3.94939 & -1.61598 \\
-1.46081 & 2.66051 & -1.68094 & 0.68022 \\
1.12651 & -1.62192 & 1.24735 & -0.55781
\end{bmatrix}
$$

$$
\Delta_1 = 10^{-2} \begin{bmatrix}
2.81318 & -5.00467 & 3.46926 & -0.84798 \\
-5.29980 & 9.24831 & -6.29206 & 2.80791 \\
4.95232 & -8.39641 & 5.73329 & -1.62170 \\
0.72029 & -1.34272 & 0.95941 & 0.54827
\end{bmatrix}
$$

$$
\Delta_2 = 10^{-2} \begin{bmatrix}
-0.69409 & 1.54874 & -0.94779 & 0.39116 \\
3.93785 & -6.79910 & 4.66564 & -1.96344 \\
-2.37995 & 4.20737 & -2.75482 & 0.95329 \\
0.70545 & -0.90615 & 0.73168 & -0.55633
\end{bmatrix}
$$

$$
\Delta_3 = 10^{-2} \begin{bmatrix}
1.67681 & -2.69078 & 1.98218 & -0.33567 \\
-0.59397 & 1.11289 & -0.71981 & 0.43504 \\
1.82472 & -2.93685 & 2.11591 & -0.43417 \\
1.28875 & -2.01749 & 1.51782 & -0.09016
\end{bmatrix}
$$

with $\begin{bmatrix} b_{11} & b_{12} & b_{13} \end{bmatrix} = \begin{bmatrix} b_{31} & b_{32} & b_{33} \end{bmatrix}$

$$
\begin{bmatrix} b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} -1.81600 & 1.23756 & -0.31382 \\
-1.81611 & 1.23775 & -0.31391
\end{bmatrix}
$$

The above data can be realized by a minimal state-space model in (3) as

$$
A_2 = \begin{bmatrix}
0.00000 & -0.19089 & 0.29060 \\
0.74393 & -86.40470 & 133.71075 \\
-0.27211 & -57.01643 & 88.22081
\end{bmatrix}
$$
Using (9) and (12) with truncation \((0, 0, 0) \leq (i, j, k) \leq (100, 100, 100)\) to evaluate the Gramians \(M_A, W_B, K_C\) and \(K\), we obtained the coordinate transformation matrix \(T\) as

\[
T = 10^3 \text{diag}\{0.01077, 2.58588, 1.68384\}
\]

which yields

\[
K = \begin{bmatrix}
1.0000 & -0.84067 & -0.83915 \\
-0.84067 & 1.0000 & 0.99999 \\
-0.83915 & 0.99999 & 1.0000 \\
\end{bmatrix}
\]

\[
K_C = 10 \begin{bmatrix}
5.70413 & -4.79529 & -4.78664 \\
-4.79529 & 5.70413 & 5.70410 \\
-4.78664 & 5.70410 & 5.70413 \\
\end{bmatrix}
\]

\[
W_B = 10^8 \begin{bmatrix}
0.02404 & -0.02399 & -0.02399 \\
-0.02399 & 0.02404 & 0.02404 \\
-0.02399 & 0.02404 & 0.02404 \\
\end{bmatrix}
\]

\[
M_A = 10^7 \begin{bmatrix}
0.03478 & 5.83540 & -5.82400 \\
-0.03479 & -5.82400 & 5.81261 \\
\end{bmatrix}
\]

In this case, the \(l_2\)-sensitivity measure was computed as

\[
S = 9.87318749 \times 10^8.
\]

Choosing \(P^{(0)} = I_3\) in (19) as an initial estimate and a tolerance \(\epsilon = 10^{-8}\) in (21) as well as in the bisection method, it took the Lagrange-based algorithm 15 iterations to converge to the solution

\[
P^{opt} = \begin{bmatrix}
2.267890 & -2.297027 & -2.289396 \\
-2.297027 & 3.274871 & 3.268974 \\
-2.289396 & 3.268974 & 3.263110 \\
\end{bmatrix}
\]

or equivalently,

\[
T^{opt} = \begin{bmatrix}
0.180276 & 1.360946 & -0.619043 \\
-1.018340 & -1.083999 & 1.030932 \\
-1.020420 & -1.079492 & 1.027884 \\
\end{bmatrix}
\]

and then the \(l_2\)-sensitivity measure was computed as

\[
J(P^{opt}, \lambda) = 3.243563304 \times 10^3
\]

where \(\lambda = 8.42923 \times 10^2\). The profiles of the \(l_2\)-sensitivity measure \(J(P, \lambda)\) and the Lagrange multiplier \(\lambda\) during the first 15 iterations are shown in Fig. 1, from which it is observed that with a tolerance \(\epsilon = 10^{-8}\) the algorithm converges within 15 iterations.

V. CONCLUSION

The realization problem of a 3-D SD digital filter has been investigated so as to reduce the \(l_2\)-sensitivity under the \(l_2\)-scaling constraints. An efficient iterative technique has been developed to solve the problem. To this end, we utilize a Lagrange function as well as an efficient bisection method, and solve the constrained optimization problem directly. Our computer simulation results have demonstrated the validity and effectiveness of the proposed technique.

REFERENCES