DESIGN OF INTERPOLOLATIVE SIGMA DELTA MODULATORS VIA A SEMI-INFINITE PROGRAMMING APPROACH

Charlotte Yuk-Fan Ho¹, Bingo Wing-Kuen Ling² and Joshua D. Reiss³

¹Telephone: +44 (0)20 78827986 Fax: +44 (0)20 78827997 Email address: charlotte.ho@elec.qmul.ac.uk Department of Electronic Engineering, Queen Mary, University of London, Mile End Road, London, E1 4NS, United Kingdom.
² Telephone: +44 (0)20 78482294 Fax: +44 (0)20 78482932 Email address: wing-kuen.ling@kcl.ac.uk Department of Electronic Engineering, Division of Engineering, King’s College London, Strand, London, WC2R 2LS, United Kingdom.
³Telephone: +44 (0)20 78827982 Fax: +44 (0)20 78827997 Email address: josh.reiss@elec.qmul.ac.uk Department of Electronic Engineering, Queen Mary, University of London, Mile End Road, London, E1 4NS, United Kingdom.

Keywords: Interpolative sigma delta modulators, noise shaping, stability, semi-infinite programming, dual parameterization.

Abstract

This paper considers the design of interpolative sigma delta modulators (SDMs). The design problem is formulated as two different optimization problems. The first optimization problem is to determine the denominator coefficients. The objective of the optimization problem is to minimize the energy of the error function in the passband of the loop filter in which the error function reflects the noise output transfer function and the ripple of the input output transfer function. The constraint of the optimization problem refers to the specification of the error function defined in the frequency domain. The second optimization problem is to determine the numerator coefficients in which the cost function is to minimize the stopband ripple energy of the loop filter subject to the stability condition of the noise output and input output transfer functions. These two optimization problems are actually quadratic semi-infinite programming (SIP) problems. By employing our recently proposed dual parameterization method for solving the problems, global optimal solutions that satisfy the corresponding continuous constraint are guaranteed if the solutions exist. The advantages of this formulation are the guarantee of the stability of the noise output and input output transfer functions, applicability to design rational IIR filters without imposing specific filter structures such as Laguerre filter and Butterworth filter structures, and the avoidance of the iterative design of numerator and the denominator coefficients because the convergence of the iterative design is not guaranteed. Our simulation results show that this proposed design yields a significant improvement in the signal-to-noise ratio (SNR) compared to the existing designs.

1 Introduction

The basic operation of SDMs is to sample the input signal at much higher rate than the Nyquist frequency and then quantize via a quantizer, typically with one bit quantization [1]. Sigma delta modulation is an increasingly popular form of analog-to-digital (A/D) conversion, and found in many real engineering systems, such as in amplitude modulation communication systems [2] and cardiac acquisition systems [3]. The popularity of SDMs is mainly due to their simple, inexpensive and robust circuit implementations, as well as achieving very high SNR because of their ability to perform noise shaping [4].

SDMs are typically designed using Butterworth filter design rules [5], and optimal designs have been performed via HSpice [6], Matlab [7] and Fortran [8]. Although these designs have considered many practical issues, the solutions obtained are not the global optimal one. This is because the optimization problems involved are not convex. Genetic algorithms have also been applied to perform the optimization [9]. However, the convergence of the genetic algorithms is not guaranteed and the computational complexity of this method is very high.

Recently, optimal SDM designs based on comb filter [10] and Laguerre filter [11] structures were proposed. However, the solutions obtained are still sub-optimal one because structural constraints (such as all the poles of the Laguerre filters are constrained to be the same) are imposed on the design. Besides, the design based on the finite horizon method [12] was proposed. However, this method is only an approximation of an infinite horizon method. Although the approximation is improved as the length of window increases, the computation complexity increases. Other existing optimal designs, such as reported in [13]-[15], are obtained mainly based on the simulation framework and lack of the theoretical support.

The objective of this paper is to formulate an SDM design problem as optimization problems based on the noise shaping characteristics of SDMs, the stopband characteristics of loop filters and the stability conditions of the noise output and input output transfer functions. The optimization problems are actually quadratic SIP problems. By applying our recently
proposed dual parameterization method for solving the problems [16], global optimal solutions that satisfy the corresponding continuous constraint are guaranteed. According to the simulations, the SDM produces a higher SNR compared to the existing designs. The outline of this paper is as follows. The problem formulation is presented in Section II. The simulation results are shown in Section III. Finally, a conclusion is summarized in Section IV.

### 2 Problem formulation

Since the phase information of the loop filters is usually not important for some applications, such as audio applications [17], an IIR loop filter is preferred compared to the FIR one. For practical reasons, it is easier to realize the SDMs if all the filter coefficients are real and the transfer function of loop filters is rational, causal, and with a unit sample delay in the numerator [17], so only real, rational and causal IIR loop filters with unit sample delay are considered in this paper. Moreover, since SDMs are usually implemented via integrators [17], there are at least one DC poles in the transfer function of the loop filters. Hence, the frequency response of the loop filters is assumed to be as follows:

$$H(\omega) = \frac{e^{-j\omega} \sum_{n=0}^{M} b_ne^{j\eta_n}}{1 - e^{-j\omega} \sum_{n=0}^{N} a_ne^{-j\omega}}., \quad (1)$$

where $j = \sqrt{-1}$, $M$ and $N$ are the numbers of roots of the polynomial of $e^{-j\omega}$ in the numerator and denominator of the transfer function of the loop filter excluding the DC poles and pure delay elements, respectively, $r$ is the number of DC poles, $a_n, b_n$ for $n = 1, 2, \ldots, N$ and $m = 0, 1, \ldots, M$ are the filter coefficients. In our consideration, $a_n, b_n \in \mathbb{R}$ for $n = 1, 2, \ldots, N$ and $m = 0, 1, \ldots, M$ where $\mathbb{R}$ denotes the set of all real numbers and $N + r \geq M + 1$. The design problem is equivalent to find an appropriate set of filter coefficients $a_n$ and $b_n$. By grouping the filter coefficients in the numerator and denominator as

$$x_n = [b_0, \ldots, b_M]^T$$

and

$$y_n = [a_1, \ldots, a_N]^T., \quad (2)$$

respectively, where the superscript $^T$ denotes the transpose, and defining

$$\eta_k(\omega) = [1, e^{-j\omega}, \ldots, e^{-j\omega M}]$$

and

$$\eta_b(\omega) = [e^{-j\omega}, e^{-j2\omega}, \ldots, e^{-jN\omega}], \quad (3)$$

then

$$H(\omega) = \frac{e^{-j\omega} \eta_b(\omega)^T x_\omega}{1 - e^{-j\omega} \eta_b(\omega)^T (1 + \eta_\omega(\omega)^T x_\omega)}. \quad (4)$$

The input output and noise output transfer functions of the SDM can be expressed as

$$T_c(\omega) = \frac{H(\omega)}{1 + H(\omega)} \quad \text{and} \quad T_p(\omega) = \frac{1}{1 + H(\omega)} \quad (6)$$

respectively. Denote the passband of the loop filter as $B_p$. For SDMs having a good SNR, the magnitude of the input output transfer function should be approximately equal to 1 and that of the noise output transfer function should be approximately equal to 0 for all frequencies in the passband of the loop filter, that is, $|T_c(\omega)| \approx 1$ and $|T_p(\omega)| \approx 0 \ \forall \omega \in B_p$. It is worth noting that if $|1 + (\eta_b(\omega)^T x_\omega)| \rightarrow 0 \ \forall \omega \in B_p$, then $|T_c(\omega)| \approx 1$ and $|T_p(\omega)| \approx 0 \ \forall \omega \in B_p$. Hence, we can define the cost function as follows:

$$J_d(x_a) = \int_{B_p} |(\eta_b(\omega))^T x_\omega|^2 d\omega. \quad (7)$$

Denote $\Re(\bullet)$ and $\Im(\bullet)$ as the real and the imaginary part of the vector inside the bracket, respectively, and define the superscript $^*$ as the conjugate operator. Although $\Re(\eta_b(\omega))^T \eta_b(\omega)^T x_\omega = |(\eta_b(\omega))^T x_\omega|^2$, (8)

$$x_\omega^T [\Re(\eta_b(\omega)^T \eta_b(\omega))^T + \Im(\eta_b(\omega)^T \eta_b(\omega))^T] x_\omega = x_\omega^T (\eta_b(\omega)^T \eta_b(\omega))^T x_\omega. \quad (9)$$

Hence, it can be shown that

$$J_d(x_a) = \int_{B_p} \frac{1}{2} x_\omega^T Q_a x_\omega + b_\omega^T x_\omega + p_\omega, \quad (10)$$

where

$$Q_a = 2 \int_{B_p} [\Re(\eta_b(\omega))^T \eta_b(\omega)^T + \Im(\eta_b(\omega)^T \eta_b(\omega))^T] d\omega, \quad (11)$$

$$b_\omega = 2 \int_{B_p} \Re(\eta_b(\omega)) d\omega, \quad \quad (12)$$

$$p_\omega = \int_{B_p} d\omega \quad (13)$$

and $Q_a$ is a positive definite matrix.

Although the cost function minimizes the energy of the function $|1 + (\eta_b(\omega))^T x_\omega|^2$ over the frequencies in the passband of the loop filter, which reflects the error energy of the noise output transfer function and the ripple energy of the input output transfer function over the corresponding frequency band, there may be a serious overshoot. In this case, then the SNR of the SDM will be poor and the SDM may suffer from limit cycle behavior. In order to avoid the occurrence of these drawbacks, the constraint based on the $H_c$ norm should be imposed in the design. Let $\delta_d^2$ be the bound on $|1 + (\eta_b(\omega))^T x_\omega|^2$ over $\omega \in B_p$, then the constraint can be represented as:

$$|1 + (\eta_b(\omega))^T x_\omega|^2 \leq \delta_d^2 \ \forall \omega \in B_p, \quad (14)$$

which can further be represented as

$$\frac{1}{2} x_\omega^T A_c(\omega) x_\omega + (c_\omega(\omega))^T x_\omega + q_\omega \leq 0 \ \forall \omega \in B_p, \quad (15)$$

where

$$A_c(\omega) = 2[\Re(\eta_b(\omega)^T \eta_b(\omega))^T + \Im(\eta_b(\omega)^T \eta_b(\omega))^T] \ \forall \omega \in B_p, \quad (16)$$

$$c_\omega(\omega) = 2 \Re(\eta_b(\omega)) \ \forall \omega \in B_p, \quad (17)$$

and

$$q_\omega = 1 - \delta_d^2 \ \forall \omega \in B_p. \quad (18)$$

Clearly, the constraint function is convex in $x_\omega$ and continuously differentiable with respect to both $x_\omega$ and $\omega$ because $A_c(\omega)$ is a positive definite matrix $\forall \omega \in B_p$. As a
result, the design of the denominator filter coefficients can be formulated as the following SIP problem:

**Problem (P₁)**

$$\begin{align*}
\text{min} \quad & J_{N}(x) = \frac{1}{2} x^T Q x + b^T x + p, \\
\text{subject to} \quad & 1 \leq \frac{1}{2} x^T A(\omega)x + (c(\omega))^T x + q, \quad \forall \omega \in B, 
\end{align*}$$

(19a)

Although problem P₁ only consists of a quadratic constraint, the constraint is a continuous function and the solution is required to satisfy the constraint for all frequencies in the passband of the loop filter. This kind of optimization problem is actually a SIP problem. To tackle a SIP problem, one way is to approximate it as a finite dimensional problem through a discretization on the index set of the continuous constraint and converts the problem to a standard quadratic programming problem [18]. However, it is not guaranteed that the continuous constraint is satisfied among the discretized points. Although the difference between the exact upper bound of a discretized constraint and that of the corresponding continuous constraint vanishes as the number of grid points increases, the computational complexity increases. To solve the problem, our recently developed dual parameterization method [16] is applied and this method guarantees to obtain a global optimal solution that satisfies the continuous constraint if the solution exists.

Besides the characteristics of the noise output and input output transfer functions are captured in the design, the stability of these two transfer functions and the frequency characteristics of the loop filter should also be considered. Our objective is to minimize the ripple energy of the loop filter in the stopband subject to the stability condition of these two transfer functions. Let the desired magnitude response of the loop filter be $\tilde{H}(\omega)$. In order to have good frequency characteristics of the loop filter, we want to achieve $|\tilde{H}(\omega)| \approx \tilde{H}(\omega)$, which implies that

$$|\tilde{H}(\omega)|^2 \geq \left| \frac{1-e^{-j\omega}}{1-e^{-j\tilde{\omega}}}(\tilde{H}(\omega) + (\tilde{H}(\omega))^T x)^T \right|^2.$$  

(20)

Since $x_0$ is obtained from solving the problem P₁, $r$ is known from the design specifications, $\tilde{H}(\omega)$ is usually zero for $\forall \omega \in B$, the cost function can be formulated as

$$J_{N}(x) = \int_{r} \left| (\tilde{H}(\omega))^T x_0 \right|^2 d\omega.$$  

(21)

The stability condition of the noise output and input output transfer functions is [18]

$$\Re \left\{ e^{-j\omega}(\eta_{N}(\omega))^T x_0 + (1-e^{-j\omega})(1 + (\eta_{N}(\omega))^T x_0) \right\} \geq 0 \quad \forall \omega \in [-\pi, \pi].$$  

(22)

which is equivalent to

$$\left| \eta_{N}(\omega) \right|^2 x_0 + \Re \left\{ (1-e^{-j\omega})(1 + (\eta_{N}(\omega))^T x_0) \right\} \geq 0 \quad \forall \omega \in [-\pi, \pi],$$  

(23)

where

$$\eta_{N}(\omega) = [ \cos \omega, \cos 2\omega, \ldots, \cos (M + 1)\omega ]^T.$$  

(24)

Hence, the optimization problem can be represented as the following SIP problem:

**Problem (P₂)**

$$\begin{align*}
\text{min} \quad & J_{N}(x) = \frac{1}{2} x^T Q x, \\
\text{subject to} \quad & A(\omega)x + c(\omega)^T x + p, \quad \forall \omega \in [-\pi, \pi], 
\end{align*}$$

(25a)

subject to $A(\omega)x + c(\omega)^T x + p, \quad \forall \omega \in [-\pi, \pi].$

(25b)

where

$$Q = 2 \left[ \Re \eta_{N}(\omega) \right] + \Im \eta_{N}(\omega) \Im \eta_{N}(\omega) \} \delta \omega.$$  

(26)

$$A(\omega) = -(\eta_{N}(\omega))^T \quad \forall \omega \in [-\pi, \pi].$$  

(27)

$$c(\omega) = -\Re \left\{ (1-e^{-j\omega})(1 + (\eta_{N}(\omega))^T x_0) \right\} \quad \forall \omega \in [-\pi, \pi].$$  

(28)

It is worth noting that problem P₁ does not depend on the numerator coefficients, so the global optimal solution of problem P₁ can be obtained via our recently proposed dual parameterization method [16]. To solve problem P₂, since the denominator coefficients are obtained from solving problem P₁, the global optimal solution of problem P₂ can then be obtained similarly. One of the advantages of this formulation is that the iterative design of the numerator and denominator coefficients is avoided because the convergence of the iterative design is not guaranteed [18].

3 Simulation results

To compare our design with the most common one via the Matlab sigma-delta toolbox [19], a fifth order SDM with a DC pole, a pure delay in the numerator of the transfer function and an oversampling ratio of 64 is considered, that is, $M = 5, N = 4, r = 1, \quad B_r = [-\pi, \pi]$ and

$$B_s = [-\pi, \pi] 
\begin{bmatrix}
\frac{\pi}{64} & \frac{\pi}{64}
\end{bmatrix}.$$  

This is because the above configuration of fifth order SDM is enough for most applications, such as audio applications [17]. By selecting $\delta_p = 9.8315 \times 10^{-2}$, the optimal SDM design problem can be formulated as SIP problems as discussed in Section II and these problems can be solved via our recently proposed dual parameterization method [16]. The magnitude responses of $|1 + (\eta_{N}(\omega))^T x_0|^2$ of both our design and the one designed via Matlab sigma-delta toolbox [19] are plotted in Figure 1. For the design using Matlab sigma-delta toolbox [19], it can be seen from Figure 1 that the magnitude response does not satisfy the constraint, while our design satisfies the required specification.

Figure 2 shows that the SNRs (calculated using the Matlab sigma-delta toolbox [19]) of both our design and that designed via the Matlab sigma-delta toolbox [19]. It can be seen from Figure 2 that our design achieves an average 5–10dB improvement compared to that of [19].

It can be verified that the magnitude of the poles of input output and noise output transfer functions of our design are 0.9928, 0.9928, 0.8556, 0.8556 and 0.6143, respectively, which all are strictly inside the unit circle. Hence, the input output and noise output transfer functions are strictly stable. Figure 3 plots the response of the output of the loop filter via our proposed design, which shows that the SDM we designed is stable.

4 Conclusion
In this paper, we have formulated SDM design problems as SIP problems and solve the problems via our recently proposed dual parameterization approach. The advantages of this formulation are the guarantee of the stability of the noise output and input output transfer functions if the solutions exist, applicability to design rational IIR filters without imposing specific filter structures such as Laguerre filter and Butterworth filter structures, and the avoidance of the iterative design of numerator and the denominator coefficients because the convergence of the iterative design is not guaranteed. Our simulation results show that the proposed design yields a significant improvement in the SNR compared to the existing designs.

Acknowledgements

The work obtained in this paper was supported by a research grant from Queen Mary, University of London.

References

Figure 1. Magnitude responses of $\left| 1 + \left( \eta_{\omega}(\omega) \right)^2 \right|^2$ of both our design and the one designed via the Matlab sigma-delta toolbox [19].
Figure 2. SNRs of both our design and the one designed via the Matlab sigma-delta toolbox [19].

Figure 3. Output response of loop filter designed via our proposed approach.