Audio Engineering Society

Convention Paper 6697

Presented at the 120th Convention
2006 May 20–23 Paris, France

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Using SIP techniques to verify the trade-off between SNR and information capacity of a sigma delta modulator

Charlotte Yuk-Fan Ho[1], Joshua D. Reiss[2] and Bingo Wing-Kuen Ling[3]

[1] Department of Electronic Engineering, Queen Mary, University of London, Mile End Road, London, E1 4NS, United Kingdom.
c.ho@qmul.ac.uk

[2] Department of Electronic Engineering, Queen Mary, University of London, Mile End Road, London, E1 4NS, United Kingdom.
Josh.reiss@qmul.ac.uk

[3] Department of Electronic Engineering, King's College London, Strand, London, WC2R 2LS, United Kingdom.
wing-kuen.ling@kcl.ac.uk

ABSTRACT

The Gerzon-Craven noise shaping theorem states that the ideal information capacity of a sigma delta modulator design is achieved if and only if the noise transfer function (NTF) is minimal phase. In this paper, it is found that there is a trade-off between the signal-to-noise ratio (SNR) and the information capacity of the noise shaped channel. In order to verify this result, loop filters satisfying and not satisfying the minimal phase condition of the NTF are designed via semi-infinite programming (SIP) techniques and solved using dual parameterization. Numerical simulation results show that the design with a minimal phase NTF achieves near the ideal information capacity of the noise shaped channel, but the SNR is low. On the other hand, the design with a non-minimal phase NTF achieves a positive value of the information capacity of the noise shaped channel, but the SNR is high. Results are also provided which compare the SIP design technique with Butterworth and Chebyshev structures and ideal theoretical SDMs, and evaluate the performance in terms of SNR and a variety of information theoretic measures which capture noise shaping qualities.

1. INTRODUCTION

It is well known that a continuous-time signal can be sampled into a discrete-time signal and achieved perfect reconstruction from the corresponding discrete-time signal via an ideal lowpass filtering if the sampling rate is higher than twice of the bandwidth of the corresponding continuous-time signal. This sampling rate is called the Nyquist sampling rate. If the sampling rate is higher than the Nyquist sampling rate, it is called an oversampling. In this case, the signal is concentrated at a very narrow band and the bandwidth of the signal is inversely proportional to the oversampling ratio (OSR). If quantization is applied on the corresponding discrete-time signal and assuming that the quantization noise is evenly distributed in the whole frequency band, then the SNR can be improved by increasing the OSR and the number of bits of the quantizer via an ideal lowpass filtering of the corresponding quantized discrete-time signal. The SNR can be further improved by applying noise shaping techniques via proper design of a loop filter. Because of the oversampling and noise shaping techniques, sigma delta modulators (SDMs) can achieve a very high SNR even for very coarse quantization steps[1]. As a result, SDMs with good performance dominate the high resolution, low frequency end of the A/D and D/A converter market[2]. They are particularly well-suited for audio applications, where their low cost and high performance for input signals below 50kHz makes them especially appealing.

The most common loop filter design methods are based on Chebyshev structures [3] or Butterworth structures[4]. However, in order to achieve good SNR, the NTF should be close to zero and the STF should be close to one in the signal band, and the NTF should be close to one and the STF should be close to zero in the noise band. Since these characteristics are defined in the frequency domain, continuous constraints should be captured in the design. To deal with this, we have recently formulated the design problem as a semi-infinite programming problem[5, 6] and solved the problem via dual parameterization[7]. Numerical simulation results show that the SDM designed using the SIP approach can achieve very high SNR.

However, there is another index for evaluating the performance of SDMs, which is based on the information capacity of the noise shaped channel. It was reported in the Gerzon-Craven noise shaping theorem[8] that maximum information capacity of the noise shaped channel is achieved if and only if the NTF is minimal phase. Hence, to achieve good performance based on the information capacity of the noise shaped channel, the NTF should be minimal phase. In this paper, loop filters satisfying the minimal phase condition of the NTF are designed. The goal of this paper is to clarify the SIP design technique and how it may be used for minimum phase NTF design, and to evaluate its performance and compare it with other design methods and with theoretical limits. The comparison between minimum phase and non-minimum phase designs allows us to demonstrate and analyze the trade-off between information capacity of an SDM and its SNR performance.

In Section 2 of this paper, we review the formulation of SDM design by Semi-Infinite Programming, and its solution using dual paramererization is presented in Section 3. Though, this

In Section 2 of this paper, the design of the loop filter satisfying the minimal phase condition of the NTF is formulated as an SIP problem. The SIP problem is solved via dual parameterization and the implementation issues are reviewed in Section 3. Though Section 3 represents a review of the state of the art, it should be noted that [6] describes the formulation of the SIP problem and the justification for dual parameterization but not its implementation. The implementation of dual parameterization was described in [7] but here the SIP problem is rephrased in the context of this paper and emphasis is placed on issues concerning practical implementation of dual paramererization as applied to SDM design.

Section 4 is focused on evaluation of SDM designs and comparison with theory. In Section 4.1, various designs are compared in terms of their SNR performance. In Section 4.2, information theoretic measures are devised and implemented which attempt to capture the effectiveness of noise shaping, and how closely the SDMs perform to theoretical limits as suggested by the Gerzon-Craven noise shaping theorem. These performance indexes include the ratio of the total power of the shaped quantization noise to that of the unshaped quantization noise, and the measure of the total loss of channel information capacity after noise shaping. Section 4.3 discusses the trade-off between the SNR and the information capacity of the noise shaped channel. A justification for this trade-off is made, and numerical simulation results are illustrated to show that there is a trade-off between the SNR and the information capacity of the noise shaped channel. Finally, in Section 5, the results are summarized and comment on future directions of research.

2. FORMULATION OF SIP-BASED DESIGN OF AN SDM WITH MINIMAL PHASE NTF

The formulation of the design of an SDM as an SIP problem was shown in [5, 6]. However, in [5, 6], the design was based on minimizing the ripple energy of NTF and STF in the signal band, subject to the H_∞ constraints on these ripples as well as the stability condition of both the STF and NTF. As a result, this design achieves very high SNR. However, the information capacity of the noise shaped channel was not considered.

The design of the loop filter is divided into two parts. The first part is to formulate the design of denominator coefficients of the loop filter as a continuous constrained optimization problem where the objective is to minimize the energy of the denominator transfer function (excluding the DC poles) in the signal band, subject to the minimal phase condition of the NTF. The second part is to formulate the design of numerator coefficients as a continuous constrained optimization problem where the objective is to minimize the energy of the numerator transfer function (excluding the delay elements) in the noise band, subject to the continuous constraint defined by the stability condition of both the NTF and the STF. The reasons for choosing these cost functions and constraints will be discussed in the following paragraphs.

We assume that the loop filter is a rational and causal filter with a unit sample delay in the numerator transfer function and there may be some DC poles in the denominator transfer function, that is:

$$\mathrm{Re}(1+(\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a) \geq 0 \quad \forall \omega \in [-\pi, \pi], \quad (1)$$

where M and N are the numbers of roots of the polynomial of $e^{-j\omega}$ in the numerator and denominator transfer functions of the loop filter (excluding the DC poles and pure delay elements), respectively, r is the number of DC poles (possibly zero), and a_n, b_m for $n = 1, 2, \cdots, N$ and $m = 0, 1, \cdots, M$ are the filter coefficients. By grouping the filter coefficients in the numerator and denominator as $\mathbf{x}_b \equiv [b_0, \ \cdots, \ b_M]^T$ and $\mathbf{x}_a \equiv [a_1, \ \cdots, \ a_N]^T$, respectively, where the superscript ^{T} denotes the transpose operator, and defining $\boldsymbol{\eta}_M(\omega) \equiv [1, \ e^{-j\omega}, \ \cdots, \ e^{-jM\omega}]^T$ and $\boldsymbol{\eta}_N(\omega) \equiv [e^{-j\omega}, \ e^{-j2\omega}, \ \cdots, \ e^{-jN\omega}]^T$, then

$$H(\omega) = \frac{e^{-j\omega}(\boldsymbol{\eta}_M(\omega))^T \mathbf{x}_b}{(1-e^{-j\omega})^r (1+(\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a)}. \quad (2)$$

The signal transfer function and noise transfer function of the SDM can be expressed as

$$STF(\omega) = \frac{H(\omega)}{1+H(\omega)}, \quad NTF(\omega) = \frac{1}{1+H(\omega)} \quad (3)$$

respectively.

2.1. Determination of denominator coefficients

Denote the passband of the loop filter, or the signal band, as

$$B_P = \left[\frac{-\pi}{OSR}; \frac{\pi}{OSR}\right], \quad (4)$$

where OSR is the oversampling ratio. For SDMs having a good SNR, the *magnitude* of the STF should be approximately equal to 1 and that of the NTF should be approximately equal to 0 for all frequencies in the signal band. This holds if

$$|1+(\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a| \to 0 \quad \forall \omega \in B_P \quad (5)$$

Hence, we define a cost function as

$$\int_{B_P} |1+(\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a|^2 \, d\omega = \frac{1}{2}\mathbf{x}_a{}^T \mathbf{Q}_a \mathbf{x}_a + \mathbf{b}_a{}^T \mathbf{x}_a + p_a, \quad (6)$$

where

$$\mathbf{Q}_a \equiv 2\int_{B_P} (\mathrm{Re}(\boldsymbol{\eta}_N(\omega))\,\mathrm{Re}(\boldsymbol{\eta}_N(\omega))^T + \mathrm{Im}(\boldsymbol{\eta}_N(\omega))\,\mathrm{Im}(\boldsymbol{\eta}_N(\omega)))d\omega$$
$$\mathbf{b}_a \equiv 2\int_{B_P} \mathrm{Re}(\boldsymbol{\eta}_N(\omega))d\omega \quad (7)$$

and

$$p_a \equiv \int_{B_P} d\omega, \quad (8)$$

in which $\mathbf{Q}_a$ is a positive definite matrix.

To capture the minimal phase condition of the NTF in the design, $H(\omega)$ should be stable. This implies that

$$\mathrm{Re}(1+(\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a) \geq 0 \quad \forall \omega \in [-\pi, \pi]. \quad (9)$$

Denote $\mathbf{A}_a(\omega) \equiv -\mathrm{Re}((\boldsymbol{\eta}_N(\omega))^T)$ and $\mathbf{c}_a(\omega) \equiv -1$. Then the design of the denominator coefficients can be formulated as the following SIP problem:

Problem (P$_1$)

$$\min_{\mathbf{x}_a} \frac{1}{2}\mathbf{x}_a{}^T \mathbf{Q}_a \mathbf{x}_a + \mathbf{b}_a{}^T \mathbf{x}_a + p_a$$
$$\text{subject to } \mathbf{A}_a(\omega)\mathbf{x}_a + \mathbf{c}_a(\omega) \leq 0 \quad \forall \omega \in [-\pi, \pi] \quad (10)$$

The SIP problem can be solved by dual parameterization[7], which guarantees the global optimal solution and satisfies the continuous constraint.

2.2. Determination of numerator coefficients

Though the characteristics of the NTF and STF for frequencies in the signal band are captured in the design, the corresponding characteristics in the noise band also need to be captured in the design. The stability of these two transfer functions and the frequency characteristics of the loop filter should be considered as well.

For SDMs having a good SNR, the *magnitude* of the STF should be approximately equal to 0 for all frequencies in the noise band. This implies that the ripple energy of the loop filter in the noise band should be small. However, $\mathbf{x}_a$ is obtained from solving the problem $\mathbf{P}_1$ and r is known from the design specifications, so the denominator transfer function does not need to be considered here. Thus, to achieve this goal, the ripple energy of the numerator transfer function should be small. The objective of the optimization problem is to minimize the ripple energy of the numerator transfer function in the noise band subject to the stability condition of the NTF and STF. The cost function can be formulated as:

$$\int_{B_S} \left| (\boldsymbol{\eta}_M(\omega))^T \mathbf{x}_b \right|^2 d\omega \cdot$$

The stability condition of the NTF and STF is

$$\mathrm{Re}\left(e^{-j\omega} (\boldsymbol{\eta}_M(\omega))^T \mathbf{x}_b + \left(1 - e^{-j\omega}\right)^r \left(1 + (\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a \right)\right) \geq 0 \quad \forall \omega \in [-\pi, \pi],$$

which is equivalent to

$$(\boldsymbol{\eta}'_M(\omega))^T \mathbf{x}_b + \mathrm{Re}\left(\left(1 - e^{-j\omega}\right)^r \left(1 + (\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a \right)\right) \geq 0 \quad \forall \omega \in [-\pi, \pi],$$

where

$$\boldsymbol{\eta}'_M(\omega) \equiv [\cos\omega, \quad \cos 2\omega, \quad \cdots, \quad \cos(M+1)\omega]^T .$$

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

Problem (P$_2$)

$$\min_{\mathbf{x}_b} \frac{1}{2} \mathbf{x}_b{}^T \mathbf{Q}_b \mathbf{x}_b \quad \text{subject to } \mathbf{A}_b(\omega)\mathbf{x}_b + \mathbf{c}_b(\omega) \leq 0, \quad \forall \omega \in [-\pi, \pi] \tag{11}$$

where

$$\mathbf{Q}_b \equiv 2 \int_{B_S} \left(\mathrm{Re}(\boldsymbol{\eta}_M(\omega)) \mathrm{Re}(\boldsymbol{\eta}_M(\omega))^T + \mathrm{Im}(\boldsymbol{\eta}_M(\omega)) \mathrm{Im}(\boldsymbol{\eta}_M(\omega))^T \right) d\omega$$

$$\mathbf{A}_b(\omega) \equiv -(\boldsymbol{\eta}'_M(\omega))^T ,$$

and

$$\mathbf{c}_b(\omega) \equiv -\mathrm{Re}\left(\left(1 - e^{-j\omega}\right)^r \left(1 + (\boldsymbol{\eta}_N(\omega))^T \mathbf{x}_a \right)\right).$$

Note that problem P$_1$ does not depend on the numerator coefficients. Thus the global optimal solution of problem P$_1$ can be obtained via dual parameterization method[7]. The denominator coefficients are obtained from solving problem P$_1$, and thus the global optimal solution of problem P$_2$ can then be obtained independently. Hence, in this formulation, iterative design of the numerator and denominator coefficients is avoided.

3. IMPLEMENTATION OF DUAL PARAMETERIZATION FOR SOLVING SIP PROBLEMS

There are many existing methods for solving SIP problems. For discretization methods[9], the continuous constraints are discretized, resulting in a finite number of convex and quadratic constraints after discretization. The problem then becomes a quadratic programming problem and can be solved efficiently via many existing solvers, such as Matlab, etc. However, this method does not guarantee that the solution obtained would satisfy the corresponding continuous constraints.

In this section, we review the dual parameterization method for solving SIP problems[7]. Since the original problem consists of continuous constraints, which are difficult to solve, we transform the original problem into an equivalent finite dimensional nonlinear programming problem via a sequence of regular convex programs and solve the finite dimensional nonlinear programming problem via the dual parameterization method. This method is guaranteed to obtain a globally optimal solution that satisfies the continuous constraint if a solution exists. For the details, please refer to [7] and [10].

Based on this theory, the corresponding finite dimensional dual problems of P$_1$ and P$_2$ are:

Problem (PDP^1)

$$\begin{aligned} \min_{\mathbf{x}_a, \boldsymbol{\mu}, \boldsymbol{\tau}} \quad & \frac{1}{2} \mathbf{x}_a{}^T \mathbf{Q}_a \mathbf{x}_a - \sum_{i=1}^{k} \left(\mathbf{c}_a(\omega_i) \right)^T \boldsymbol{\mu}_i \\ \text{subject to} \quad & \mathbf{Q}_a \mathbf{x}_a + \mathbf{b}_a + \sum_{i=1}^{k} \left(\mathbf{A}_a(\omega_i) \right)^T \boldsymbol{\mu}_i = \mathbf{0} \\ & \boldsymbol{\mu}_i \geq \mathbf{0}, \omega_i \in \Delta \quad \text{for } i = 1, 2, \cdots, k \end{aligned} \tag{12}$$

and

Problem (PDP^2)

$$\min_{\mathbf{x}_b,\boldsymbol{\mu},\boldsymbol{\tau}} \frac{1}{2}\mathbf{x}_b{}^T\mathbf{Q}_b\mathbf{x}_b - \sum_{i=1}^{k}\left(\mathbf{c}_b\left(\omega_i\right)\right)^T \boldsymbol{\mu}_i$$
$$\text{subject to} \quad \mathbf{Q}_b\mathbf{x} + \sum_{i=1}^{k}\left(\mathbf{A}_b\left(\omega_i\right)\right)^T \boldsymbol{\mu}_i = \mathbf{0} \quad , \quad (13)$$
$$\boldsymbol{\mu}_i \geq \mathbf{0}, \omega_i \in \Delta \;\; \text{for } i = 1,2,\cdots,k$$

where $\boldsymbol{\mu}_i$ and ω_i are, respectively, the discrete multipliers and the discrete frequencies, in which $\boldsymbol{\tau} \equiv [\omega_1, \omega_2, \cdots, \omega_k]$and k is the dimension of $\mathbf{x}$.

The implementation procedures [7] of the dual parameterization method are summarized in the following algorithm. First we initialize parameters in the algorithm (Step 1), then compute a local optimal solution by solving a finite dimensional nonlinear programming problem (Steps 2-4). Finally, the global optimal solution is computed via a local search for the finite dual problem (Step 5).

Algorithm

Step 1. *Initialization*

Arbitrarily choose an initial guess of the filter coefficients $\mathbf{x}^0 \in \mathbb{R}^S$, where S is the number of filter coefficients to be determined. For the problem P$_1$, $S = N$, while for the problem P$_2$, $S = M + 1$.

Choose a small positive number $\varepsilon > 0$ which defines the acceptable error on the constraints.

Choose a minimum iteration number N' to prevent the algorithm from terminating prematurely.

Choose a sequence of initial finite parameterization sets

$$\Delta_i = \left\{\omega_j^i : j = 1,2,\cdots,k_i\right\}$$

that satisfies the metric

$$d(\Delta_i, \Delta) \equiv \max_{\varpi\in\Delta} \min_{\omega\in\Delta_i} |\varpi - \omega| \to 0 \,,$$

where ϖ and ω are frequencies.

Set the temporarily finite parameterization sets E_0 to the empty set, and the iteration index $i = 0$.

Step 2. *Determine the set of discrete frequencies*

Increment the iteration index by 1, that is, $i = i + 1$. Find a point in the initial finite parameterization sets $\varpi_i \in \Delta_i$ such that the constraints reach their maximum value at ϖ_i for all the points in Δ_i, that is

$$g_{\max}\left(\mathbf{x}^{i-1}, \varpi_i\right) = \max_{\omega\in\Delta_i} g_{\max}\left(\mathbf{x}^{i-1}, \omega\right),$$

where $g_{\max}(\mathbf{x}, \varpi_i)$ is the maximum element among all the elements in the constraint vector $\mathbf{g}(\mathbf{x}, \varpi_i)$.

If the constraints meet the specifications at ϖ_i, that is,

$$g_{\max}\left(\mathbf{x}^{i-1}, \varpi_i\right) < \varepsilon \,,$$

then set the finite parameterization sets Z_i to the previous temporarily finite parameterization sets, that is

$$Z_i = E_{i-1} \,.$$

If the iteration index reaches the minimum iteration number N', that is, $i \leq N'$, then go to Step 5. Otherwise, set the current values of the filter coefficients $\mathbf{x}^i$, the multipliers $\boldsymbol{\mu}^i$ and the temporarily finite parameterization sets E_i to their previous values, that is

$$\left(\mathbf{x}^i, \boldsymbol{\mu}^i\right) = \left(\mathbf{x}^{i-1}, \boldsymbol{\mu}^{i-1}\right)$$

and

$$E_i = E_{i-1} \,,$$

and repeat Step 2 again.

Otherwise, set ϖ_i and the previous temporarily finite parameterization sets as the current finite parameterization sets , that is,

$$Z_i = E_{i-1} \cup \{\varpi_i\}.$$

Step 3. *Solving the problem* $\mathrm{PDP}(Z_i)$

Solve the problem $\mathrm{PDP}(Z_i)$ to obtain a solution for $\left(\mathbf{x}^i, \boldsymbol{\mu}^i\right)$, where the problem $\mathrm{PDP}(Z_i)$ is the finite dimensional problem of **PDP** subject to the finite parameterization sets Z_i, that is $Z_i = \{\omega_i\}$.

Step 4. Set the current temporarily finite parameterization sets as a subset of the current finite parameterization sets, that is $E_i \subset Z_i$, with no more than $S + 1$ points such that the solution of the problem $\mathrm{PDP}(E_i)$ is in the form $\left(\mathbf{x}^i, \boldsymbol{\mu}^i\right)$. Then go to Step 2 again.

Step 5. *Local search for the finite dual problem*

Suppose the current finite parameterization sets Z_i has k points $\varpi_1, \varpi_2, \cdots, \varpi_k$. Starting from $(\mathbf{x}^i, \boldsymbol{\mu}^i, \boldsymbol{\tau}^i)$, where $\mathbf{x}^i$ and $\boldsymbol{\mu}^i$ are defined previously and $\boldsymbol{\tau}^i = [\varpi_1, \varpi_2, \cdots, \varpi_k]$ is the k tuple formed by the points in Z_i, find a local minimum $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*)$ for the problem PDP_k, where the problem PDP_k is the discretization version subject to the initial parameterization sets Δ_i. Then $\mathbf{x}^*$ is taken as the solution for the problem P.

Solving the problem $\mathrm{PDP}(Z_i)$ in Step 3 and finding the local minimum in Step 5 can be accomplished by using existing optimization solvers. The above algorithm can be summarized by the flowchart depicted in Figure 1.

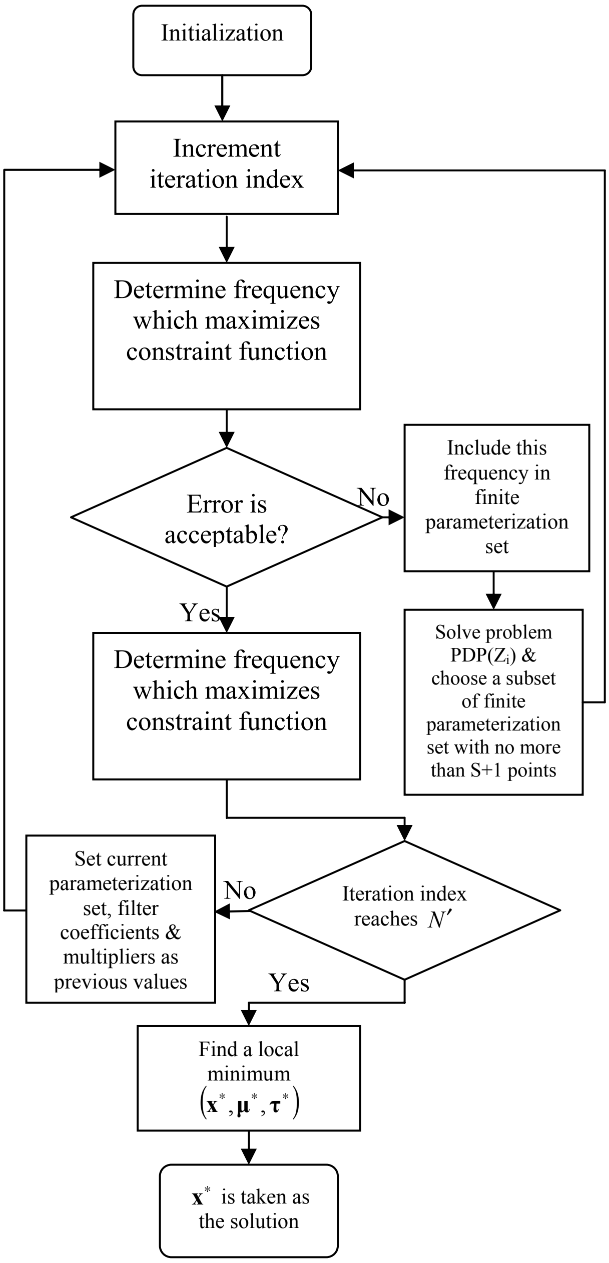
Initialization
Increment iteration index
Determine frequency which maximizes constraint function
Error is acceptable?
No
Include this frequency in finite parameterization set
Yes
Determine frequency which maximizes constraint function
Solve problem PDP(Z_i) & choose a subset of finite parameterization set with no more than S+1 points
Set current parameterization set, filter coefficients & multipliers as previous values
No
Iteration index reaches N'
Yes
Find a local minimum $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*)$
$\mathbf{x}^*$ is taken as the solution

Figure 1. Flowchart showing the implementation of the dual parameterization method.

4. PERFORMANCE EVALUATION

To evaluate the performance of SDMs designed using Semi-Infinite Programming, as well as the trade-off between the SNR and the information capacity of noise shaped channel, SDMs were designed using SIP approaches with and without minimal phase NTF, and using Butterworth and Chebyshev filter design techniques. Design of Chebyshev structures[3] was accomplished via the function “synthesisNTF” in the delta-sigma matlab toolbox[11], and design of Butterworth structures[4] was accomplished via imposing the maximally flat condition on the design. Where possible, designs were also compared with theoretical limits.

Each SDM, unless otherwise noted, had an oversampling ratio *OSR*=64 (sampling frequency 64x44.1kHz), one DC pole, *r*=1, zero initial conditions, and the number of roots in the numerator and denominator transfer functions are both 4, *M*=*N*=4 (see Eq. (2)), i.e. filter order 4+1=5. The quantizer was single bit with the decision boundary at zero and the saturation level at one.

4.1. SNR-based Performance Evaluation

First, we compare the performances of each SDM design in terms of its signal-to-noise ratio. The theoretical limit of the signal-to-noise ratio may be estimated by [1]:

$$\begin{aligned} SNR_{estimated} &= 10\log_{10}\left(\sigma_x^2\right) - 10\log_{10}\left(\sigma_e^2\right) \\ &-10\log_{10}\left(\frac{\pi^{2N}}{2N+1}\right) + (20N+10)\log_{10} R \end{aligned} \qquad (14)$$

Here, N represents the filter order, σ_x^2 and σ_e^2 are, respectively, the power of the input signal and the power of the quantization noise. If a sinusoidal input with magnitude U is employed, then $\sigma_x^2 = U^2/2$. For the quantizer with dynamical range between -1 and 1, then $\sigma_e^2 = \frac{1}{3(2^L-1)^2}$, where L is the number of bits of the quantizer.

In the case of a fifth order SDM with 64 times OSR, the theoretical limit of SNR given in (14) reduces to:

$$SNR_{estimated} = 20\log_{10} U + 20\log_{10}\left(2^L - 1\right) + 161.14 \qquad (15)$$

Figure 2 shows the relationship between the SNRs and the input magnitudes, where the initial conditions of the above SDMs are zero, the quantization region of the quantizer is between -1 and 1, $L=1$, $R=64$, and $N=5$. The theoretical limit of SNR was found from (6) and the measured SNRs were computed using the delta-sigma matlab toolbox [3, 11]. The SDM designed via the SIP approach without the minimal phase condition of the NTF consistently outperforms the SDMs designed via the Chebyshev structure, the Butterworth structure and the SIP approach with the minimal phase condition of the NTF by approximately 3.75dB, 3.04dB and 9.24dB, respectively, in their corresponding stable regions.

The non-minimal phase SDM designed via the SIP approach also has a higher dynamic range, with stable behavior up to inputs of approximately 0.68, compared to 0.66 for the Chebyshev structure and minimal phase SIP design, and 0.60 for the Butterworth structure. This is a surprising result, since more noise shaping is often thought to result in a trade-off with dynamic range, and since no attempt was made to design for stable behavior at high amplitudes.

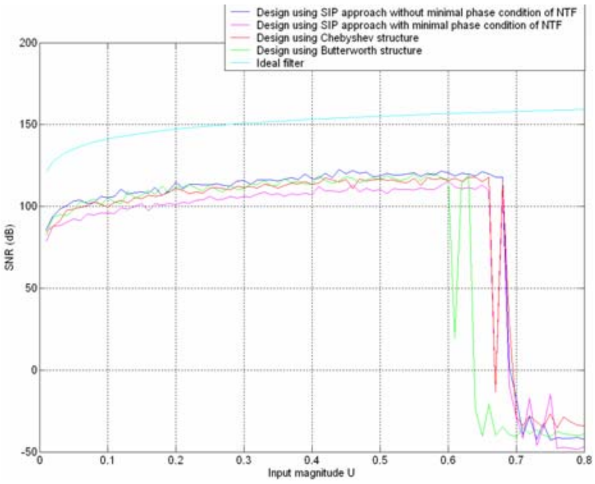

Figure 2. Relationships between SNRs and input magnitudes.

Figure 3 depicts the relationship between the SNR and the number of bits in the quantizer. Initial conditions of all SDMs are zero, the theoretical limit of SNR is given by (15), and the SNRs of the various SDMs are calculated via the delta-sigma matlab toolbox[11]. As before, a fifth order SDM with OSR=64 was used for all designs. We chose an input amplitude of U=0.44 because this guarantees stability for all the simulated designs. In almost all cases, the SDM designed via the non-minimum phase SIP design outperforms the SDMs designed via the Chebyshev structure, the Butterworth structure and the SIP approach with minimal phase NTF by 3.58dB, 1.85dB and 9.37dB, respectively. Improvements over Chebyshev and Butterworth structures are particularly noticeable for a low number of bits.

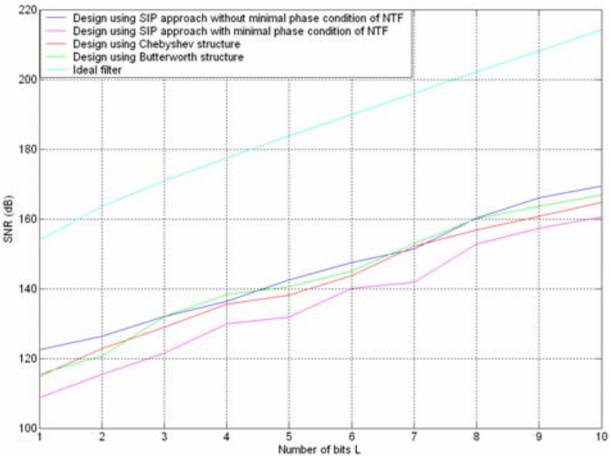

Figure 3. Relationships between SNRs and the number of bits of the quantizer.

Figure 4 shows the relationships between the SNRs and the OSRs, where the initial conditions of SDMs are zero, the quantization region of the quantizer is between -1 and 1, $L=1$, $N=5$, U was set to 0.23 in order to guarantee stability and theoretical limit of SNR were computed from (14). The non-minimal phase SIP design outperforms the optimized Chebyshev design by at least 4.48dB, and shows increased comparative performance with increased OSR.

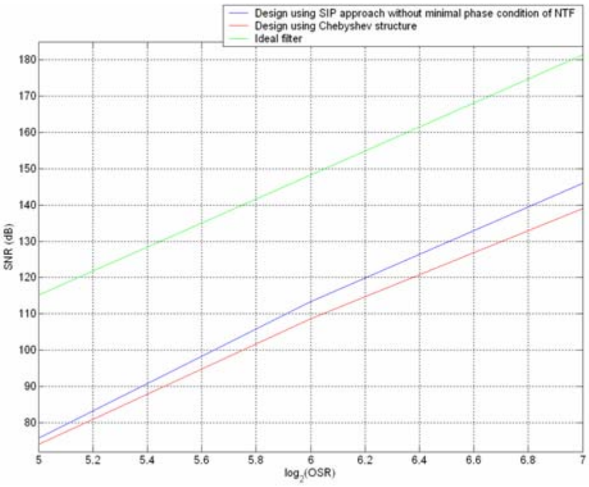

Figure 4. Relationship between SNRs and OSRs.

Finally, Figure 5 shows the relationships between the SNRs and the filter orders, where $U=0.23$, $L=1$, and OSR=64. The Chebyshev design and SIP design without the minimal phase condition of the NTF show nearly identical performance at filter orders of 3 and 7, but the SIP design without the minimal phase condition of the NTF shows up to a 4.72dB improvement when the filter order N=5. This is significant, since this type of filter order is commonly used, particularly in audio applications, because it represents a best compromise between noise shaping and stability.

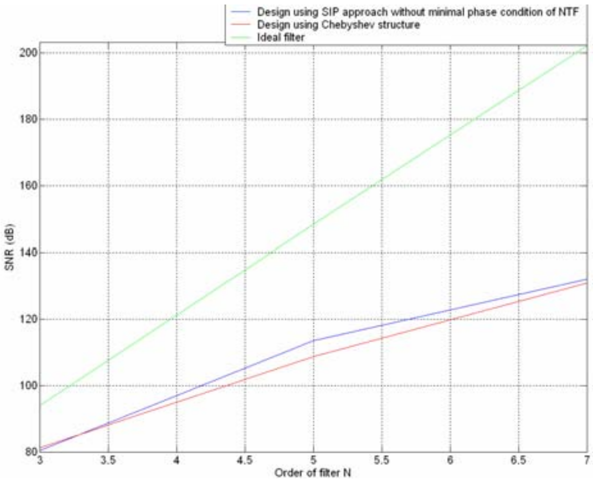

Figure 5. Relationship between SNRs and orders of filters.

Measurement of the SNR allows us to estimate the effective resolution of the design. This is given by the

$$R_{eff} = \frac{SNR - 1.76}{6.02}. \qquad (16)$$

Note that this is not quite the Effective Number of Bits, which is more accurately given by replacing SNR with the SINAD in the above formula[12]. For comparison, results for the SNR and effective resolution for the various designs are given in Table 1, where OSR=64, the filter order is 5, and a 1 bit quantiser is used.

In all results depicted in Figures 2-5, though the SDM designed using the SIP approach with a non-minimal phase NTF outperforms other methods, it does not come close to the theoretical limit of SNR. The theoretical limit of SNR given in (14) makes several assumptions, most notably the assumption of uniform distribution of the quantization error. Its assumptions also introduce significant approximations for high order SDMs. This is shown dramatically in Figure 5. The theoretical limit of SNR is 70.08dB higher than that of the SIP design with non-minimal phase NTF when N=7, and neither simulated design shows the predicted 108dB improvement from N=3 to N=7. Another difficulty in this case is the inability to estimate very high signal-to-noise ratios, where finite precision results in serious underestimates of SNRs above about 130dBs. Finally, the authors know of no reported SDM design which achieves close to the theoretical limit suggested by (14) for a high order (N>2) filter. This is partly because of the difficulties in creating ideal filters. But it is also because such designs are unstable, thus limiting the input magnitude and making the approximations concerning quantization error more inaccurate.

Nevertheless Eq. (14) is known to produce an upper bound on the SNR that can be extremely accurate for unshaped, or first and second order filters. Furthermore, the expected $(6L+1)$dB improvement with doubling the OSR and approximately 6dB improvement with adding a bit to the quantizer are observed in all designs.

4.2. Information Theoretic and Noise Shaping Metrics

In [8], Gerzon and Craven considered the sigma delta modulator as a transmission channel, and used information theoretic considerations to show that for a specified NTF, the information capacity of the noise shaped channel cannot exceed that of the non-noise shaped channel. In other words, the noise shaping filter with the smallest possible output error power is the filter that leaves the information capacity of the channel unaltered (maximum).This implies that,

$$G \equiv \frac{1}{2}\int_{-\pi}^{\pi} \log_2 \left|NTF(\omega)\right| d\omega \ge 0 \qquad (17)$$

where equality is obtained (maximized use of the information capacity of the noise shaped channel) for minimum phase filters.

This result may be rephrased as stating that the areas above and below the 0-dB line (on a decibel plot of the signal spectrum vs linear frequency) will be equal for any optimal (i.e., minimum-phase) noise shaper.

In [13], this result was used to show that, if the NTF is specified just over the signal band, then the ratio of the total power of the shaped quantization noise to that of the unshaped quantization noise is minimized by setting the NTF to a constant or flat shape over the remaining bands. Such an ideal NTF is depicted in Figure 6.

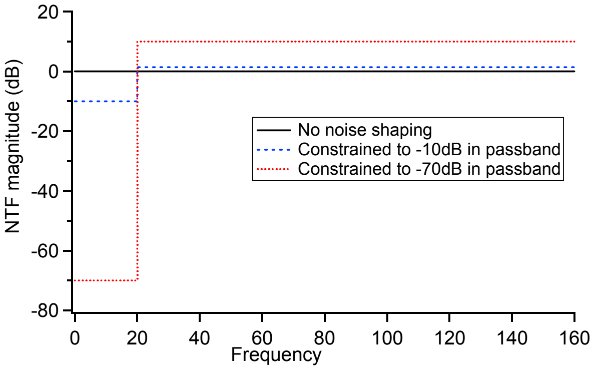

Figure 6. NTFs of SDMs with ideal filters for OSR=8.

Hence, the ideal NTFs can be expressed in the form of:

$$\left|NTF(\omega)\right| = \begin{cases} A & |\omega| \le \pi/OSR \\ B & \pi/OSR < |\omega| < \pi \end{cases}, \quad (18)$$

where $NTF(\omega) = NTF^*(-\omega)$, in which the superscript * denotes the complex conjugate. Based on the noise shaping theorem[8], $A \neq 0$ and $G = 0$ if the NTF is minimal phase.

From (17) and (18), we have that, for a minimum phase NTF,

$$\begin{aligned} G &= \int_0^{\frac{\pi}{OSR}} \log_2 A d\omega + \int_{\frac{\pi}{OSR}}^{\pi} \log_2 B d\omega \\ &= \frac{\pi}{OSR} \log_2 A + \left(\pi - \frac{\pi}{OSR}\right) \log_2 B = 0 \end{aligned} . \quad (19)$$

This implies that

$$\log_2 A = (1 - OSR) \log_2 B \Rightarrow A = B^{1-OSR}. \quad (20)$$

Eq. (20) implies that, for comparison purposes we should consider two types of ideal filter. The first allows direct comparison of the SIP design, which imposes a constrained NTF in the signal band, with the ideal filter with the same constraint. Since the loop filter designed via the SIP approach without the minimal phase condition of the NTF can achieve the bound $A = 2.3936 \times 10^{-5}$ on the NTF in the signal band, we have that ideally, $B = 1.1840$.

For the second type of ideal filter, we consider an ideal constraint on the NTF in the passband. Since the NTF should be close to zero in the signal band, while the NTF should be close to one in the noise band, $A = 0$ and $B = 1$.

4.2.1. Information Capacity Metrics

The first metric we employed for comparison is based on the information capacity of the noise shaped channel, and defined by Eq. (17). For the first ideal (non-zero) NTF, by definition $G = 0$. For the second ideal NTF, G is negative infinity because $A = 0$.

However, for a lowpass SDM, we are primarily concerned with lowpass performance. That is, we would like to measure how well the information capacity has been utilized in the signal band. Therefore, it is more practical to compute this metric *in the signal band only*, that is:

$$G' = \frac{1}{2} \int_{-\pi/OSR}^{\pi/OSR} \log_2 \left|NTF(\omega)\right| d\omega. \quad (21)$$

The value of G' is always less than zero because $|NTF(\omega)| \ll 1$ in the signal band. It is negative infinity for the SDM with the second ideal NTF because A=0.

Table 1 lists how the various SDM designs perform in terms of G and G'. The value of G for the SDM designed via an SIP approach with the minimal phase condition of NTF is still nonzero because there is a DC pole on the corresponding loop filter transfer function. For this reason, none of the designs achieve G=0, though the minimum phase design outperforms the non-minimum phase design, as expected.

The minimum phase filter actually performs poorly in terms of G'. This is most likely due to overshoot within the passband (see Figure 7) . Interestingly, on this metric, all other simulated filters actually outperform the ideal filter with non-zero NTF, with the non-minimum phase SDM performing best. This is because the ideal filter assumes a constant NTF within the signal band set to the constraint $A = 2.3936 \times 10^{-5}$, whereas the other filters often have NTFs that are mostly well-below A in the signal band.

4.2.2. Noise Shaping Metrics

It is important to evaluate the noise shaping characteristics of the loop filter. To achieve this goal, the ratio of the *total* power of the shaped quantization noise to that of the unshaped quantization noise is evaluated[13]. Assume that the unshaped quantization noise power spectral density is flat in the frequency spectrum with the magnitude denoted as S_e, then the ratio of the *total* power of the shaped quantization noise to that of the unshaped quantization noise is:

$$\begin{aligned} K &= \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_e \left|NTF(\omega)\right|^2 d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_e d\omega} \\ &= \frac{S_e \int_{-\pi}^{\pi} \left|NTF(\omega)\right|^2 d\omega}{S_e \int_{-\pi}^{\pi} d\omega} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|NTF(\omega)\right|^2 d\omega \end{aligned} . \quad (22)$$

Note that this equation is a little bit different from that in [13] because the original formula had a typing error.

For the SDM with the first ideal NTF (constrained, non-zero), we have, from Eq. (20),

$$\begin{aligned} K &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|NTF(\omega)\right|^2 d\omega \\ &= \frac{2}{2\pi}\left(\int_0^{\frac{\pi}{OSR}} A^2 d\omega + \int_{\frac{\pi}{OSR}}^{\pi} B^2 d\omega\right). \\ &= \frac{B^{2-2R}+(OSR-1)B^2}{OSR} \end{aligned} \tag{23}$$

For the SDM with the second ideal NTF (*A*=0, *B*=1), we have:

$$\begin{aligned} K &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|NTF(\omega)\right|^2 d\omega \\ &= \frac{2}{2\pi}\int_{\frac{\pi}{OSR}}^{\pi} d\omega = 1-\frac{1}{OSR} \end{aligned}. \tag{24}$$

As in Section 4.2.1, we are primarily concerned with the noise shaping characteristics in the low pass region. Thus it is more practical to compute this metric *in the signal band only*, that is:

$$\begin{aligned} K' &= \frac{\frac{OSR}{2\pi}\int_{-\pi/OSR}^{\pi/OSR} S_e \,|\, NTF(\omega)\,|^2 \, d\omega}{\frac{OSR}{2\pi}\int_{-\pi/OSR}^{\pi/OSR} S_e d\omega} \\ &= \frac{OSR}{2\pi}\int_{-\pi/OSR}^{\pi/OSR} |\, NTF(\omega)\,|^2 \, d\omega \end{aligned} \tag{25}$$

As with *G*', *K*' is always less than zero because $|NTF(\omega)| \ll 1$ in the signal band, and negative infinity (on a decibel scale) for the SDM with zero NTF in the signal band because *A*=0.

The results are shown in Table 1, where values for *K* and *K*' are given on a decibel scale. From these results, we can see that the ratio of the *total* noise power of the shaped quantization noise to that of the unshaped quantization noise is smallest for the Chebyshev design and the minimal phase SDM designed via the SIP approach. On the other hand, the Butterworth structure performs worst with the highest *total* noise power.

However, if we only consider the ratio of the power of the shaped quantization noise to that of the unshaped quantization noise *in the signal band only*, *K'*, then the SIP approach without the minimal phase performs best, while a minimal phase design performs worst.

To compare the noise shaping characteristics, the NTFs of the above SDMs are plotted in Figure 7. It can be seen that the SDM designed via the SIP approach without the minimal phase condition of the NTF has the best noise shaping characteristics. There are on average, approximately 9.5dB, 5.1dB and 2.7dB improvements over the frequency spectrum 0-20kHz compared to the SDMs designed via the Chebyshev structure, the Butterworth structure and the SIP approach with the minimal phase condition of NTF, respectively. However, for the SDM designed with the SIP approach with the minimal phase condition of the NTF, there is a serious overshoot in the NTF spectrum. This is because the H_∞ constraints on the NTF spectrum, which bound the NTF throughout the signal band, have been removed in the design procedure in order to introduce new minimal phase constraints. Hence, it accounts for the worse performances of K' and G'.

4.3. Trade-off between SNR and information capacity of noise shaped channel

The signal-to-noise ratio reflects the reconstruction error of the analog-to-digital conversion. High SNR corresponds to low reconstruction error. In order to achieve high SNR, STF should be close to one and NTF should be close to zero in the signal band. As a result, the frequency response of the loop filter should be infinity in the signal band.

However, if the NTF is minimal phase, this implies that the loop filter is stable and the region of convergence of the loop filter transfer function includes the unit circle. In which case, the frequency response of the loop filter is well defined for all frequencies, which contradicts the property of the frequency response of the loop filter having high SNR performance. As a result, there is a trade-off between the SNR and the information capacity of noise shaped channel. This agrees with the results described in Sections 4.2 and 4.3.

5. CONCLUSION

In this paper, optimal SDM designs based on semi-infinite programming, with minimal and non-minimal phase NTFs were compared with Butterworth structures, Chebyshev structures and theoretical designs. The non-minimal phase SDM designed using SIP demonstrated high SNR and high information theoretic performance in the signal band, whereas the minimal phase design had generally poor performance in the signal band. Yet this minimal phase design had high performance over the entire spectrum and came closer to the limits suggested by the Gerzon-Craven noise shaping theorem than any of the other designs. This suggests a trade-off between the SNR and the information capacity of a channel. It may be partly accounted for by the explanation given in Section 4.3, and also partly due to overshoot of the NTF in the signal band of the minimum phase design.

Further work in this area would involve performance analysis of designs without DC poles, designs incorporating the minimal phase constraint while preventing overshoot, and comparison with other optimized design methods.

SDM design methods	*SNR* (dB)	R_{eff} (bits)	*G* (bits)	*G'* (bits)	*K* (dB)	*K'* (dB)
SIP approach without minimal phase NTF	113.86	18.62	0.0050	-0.848	3.15	-98.16
SIP approach with minimal phase NTF	104.62	17.09	0.0044	-0.796	3.13	-89.50
Chebyshev structure	110.11	18.00	0.0039	-0.807	3.13	-95.56
Butterworth structure	110.82	18.12	0.0151	-0.837	3.38	-97.77
Ideal filter with nonzero NTF in signal band	156	25.02	0	-0.754	1.40	-114.36
Ideal filter with zero NTF in signal band	156	25.02	-∞	-∞	-0.0684	-∞

Table 1. Various performance indices of SDMs designed via SIP approaches, Chebyshev structure and Butterworth structure. Ideal SNRs assumed a 1-bit, 64 times oversampled, fifth order SDM.

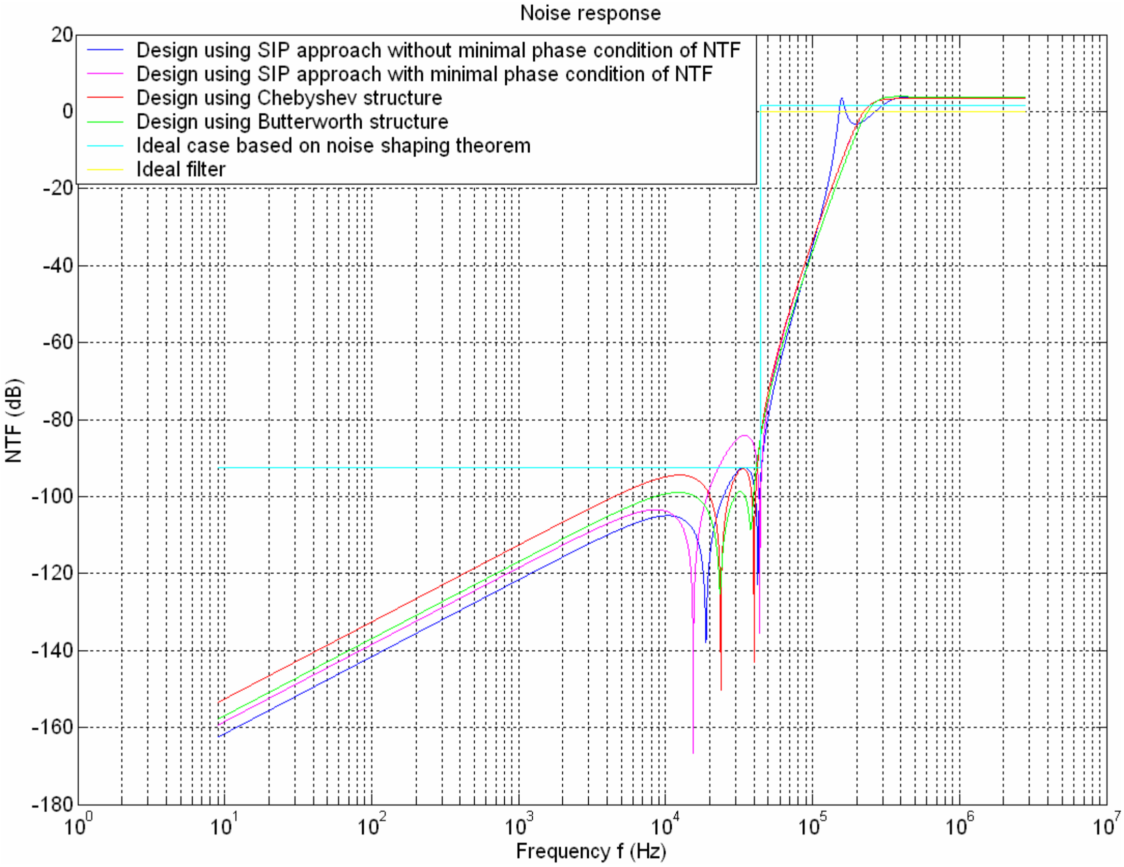

Figure 7. NTF of SDMs designed via SIP approaches, Chebyshev structure and Butterworth structure.

6. ACKNOWLEDGEMENTS

This work was supported by a research grant from Queen Mary, University of London.

7. REFERENCES

[1] P. M. Aziz, S. H. V., and J. V. Spiegel, "An overview of sigma-delta converters: how a 1-bit ADC achieves more than 16-bit resolution," *IEEE Signal Processing Magazine*, pp. 61-84, 1996.

[2] W. Kester, "Which ADC Architecture Is Right for Your Application?," *Analog Dialogue*, vol. 39, pp. 11-18, 2005.

[3] S. Norsworthy, R. Schreier, and G. Temes, *Delta-Sigma Data Converters*: IEEE Press, 1997.

[4] D. Reefman and E. Janssen, "Signal processing for Direct Stream Digital: A tutorial for digital Sigma Delta modulation and 1-bit digital audio processing," Philips Research, Eindhoven, White Paper 18 December 2002

[5] C. Y.-F. Ho, B. W.-K. Ling, J. D. Reiss, Y.-Q. Liu, and K.-L. Teo, "Design of Interpolative sigma delta modulators via semi-infinite programming," *to appear in IEEE Transactions on Signal Processing*, 2005.

[6] C. Y.-F. Ho, B. W.-K. Ling, and J. D. Reiss, "Design of Interpolative Sigma Delta Modulators via a Semi-infinite Programming Approach," Proceedings of the Advanced A/D and D/A Conversion Techniques and Their Applications (ADDA), Limerick, 2005.

[7] C. Y.-F. Ho, B. W.-K. Ling, Y. Q. Liu, P. K. S. Tam, and K. L. Teo, "Efficient algorithm for solving semi-infinite programming problems and their applications to nonuniform filter bank designs," *to appear in IEEE Transactions on Signal Processing*, 2006.

[8] M. Gerzon and P. G. Craven, "Optimal Noise Shaping and Dither of Digital Signals," Proceedings of the 87th AES Convention, New York, New York, USA, 1989.

[9] H. H. Dam, S. Nordebo, C. A., and K. L. Teo, "Frequency domain design for digital Laguerre networks," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, pp. 578-581, 2000.

[10] M. S. Bazarra and C. M. Shetty, *Nonlinear programming: theory and algorithm*. Chichester, New York: Wiley, 1979.

[11] R. Schreier, "The delta-sigma modulators toolbox version," 6.0 ed: Analog Devices Inc., 2003. www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=19

[12] T. UNIPR, ITALTEL, INFINEON technologies, ENSERB and INESC Porto, "DYNAD: Methods and draft standards for the DYNamic characterization and testing of Analogue to Digital converters." http://www.fe.up.pt/~hsm/dynad

[13] R. Nawrocki, G. J. M., and M. B. Sandler, "Information-theoretic constraints on noise shaping networks with minimum noise power gain," *International Journal of Circuit Theory and Applications*, vol. 27, pp. 437-441, 1999.