2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Optimal FIR Filters for Sigma-Delta Modulated Signals

Ingo Wiemer and Wolfgang Schwarz

Dresden University of Technology
Faculty of Electrical Engineering, IEE
Mommsenstraße 13, 01062 Dresden, Germany
Email: {wiemer, schwarz}@iee.et.tu-dresden.de

Abstract—In order to decode sigma-delta modulated sequences, linear filters with nearly rectangular passband characteristic are used in most applications. But they do not achieve optimal performance.

In this paper we propose a method to calculate the optimal coefficients of a finite impulse response (FIR) filter for sigma-delta modulators with bandlimited input signals. We will also show that it is appropriate for the filter design to model the nonlinear quantizer by an additive noise source. The results will be verified by simulations.

1. Introduction

Sigma-delta modulators (ΣΔM) use oversampling together with noise shaping to achieve high signal to quantization noise ratios (SQNR). In order to obtain a robust and simple analog circuit ΣΔMs apply a low resolution quantizer with often just one bit. Thus the modulator output is a high rate low resolution signal. It has to be converted back into a high resolution signal with a lower sampling rate. This is the task of the decoder.

The overall system of modulator and decoder is depicted in Fig. 1. The modulator can be considered as a coder generating a coded sequence $\mathbf{y}$ from the input sequence $\mathbf{x}$. $\tilde{\mathbf{x}}$ is the output of the decoder and should approximate $\mathbf{x}$ as close as possible. The quality of this approximation is characterized by the signal to quantization noise ratio (SQNR)

$$\mathrm{SQNR} = \frac{\mathrm{E}\left\{x^2(k)\right\}}{\mathrm{E}\left\{(x(k) - \tilde{x}(k))^2\right\}} , \qquad (1)$$

which is to be maximized. This is equivalent to minimizing the mean squared error (MSE)

$$\min_{\tilde{x}(k)}(\mathrm{MSE}) = \min_{\tilde{x}(k)}\left(\mathrm{E}\left\{(x(k) - \tilde{x}(k))^2\right\}\right) . \qquad (2)$$

In this paper we use linear finite impulse response (FIR) filters for the decoder part. They have a low complexity and a linear phase. There already exist methods to find the optimal filter coefficients for constant input signals [1–3]. In this paper we present a method to compute the optimal filter coefficients for bandlimited input signals.

This work was supported by the Deutsche Forschungsgemeinschaft (Project: SFB 358/E1)

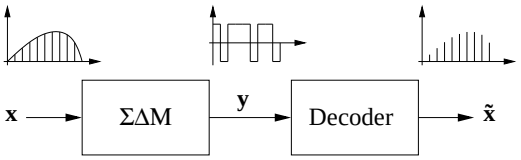

Fig. 1: Analog to digital converter employing a ΣΔM (modulator and decoder)

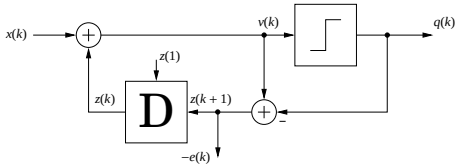

Fig. 2: SMI structure

Most literature on filter design for ΣΔMs [4] deals with the implementation of the filter, e. g. how many stages of filters are used and how these stages have to be designed to approximate a rectangular passband characteristic. Here we are interested in the optimal filter coefficients, if a linear finite impulse response (FIR) filter is used. How this filter is implemented and in how many stages will be part of the future work.

First the system equations will be derived. The parameters of the input signal and the quantization noise are assumed to be uncorrelated random variables. This makes it possible to find an equation for the MSE leading to a linear minimization problem. Thus the optimal filter coefficients can be calculated. In the results section our assumptions will be verified and the SQNR performance of our filter is presented.

2. Optimal FIR Filter and its SQNR

A simple first order modulator can be represented by the single module integrator (SMI) structure as displayed in Fig. 2. The modulator input at time instance k is $x(k)$. $q(k)$ is the output of the one bit quantizer, $z(k)$ is the modulator state and $e(k)$ is the quantization error. The initial state is $z(1) = z_0$. The quantizer function is

$$q = \begin{cases} -1 & ;\ v < 0 \\ +1 & ;\ v \geq 0 \end{cases} = \mathrm{sgn}(v) . \qquad (3)$$

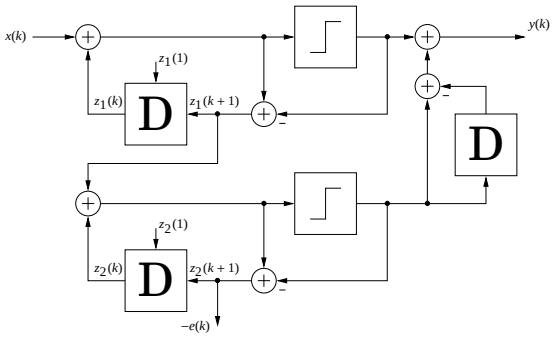

Fig. 3: Second order MASH structure

The state equation of the SMI structure is

$$z(k+1) = -e(k) = z(k) + x(k) - \operatorname{sgn}(z(k) + x(k)) \,. \quad (4)$$

The output equation is

$$y(k) = q(k) = \operatorname{sgn}(z(k) + x(k)) \,. \quad (5)$$

These equations yield the output of the SMI-structure to be

$$y(k) = x(k) + e(k) - e(k-1) \,. \quad (6)$$

Eq. (6) can be modified for any modulator type, e. g. cascades of SMI structures or multiloop modulators. Here we focus on the so called multi stage noise shaping (MASH) structure consisting of cascaded SMI structures. They are connected in a way to cancel out the quantization errors of all but the last SMI structure. A second order MASH structure is shown exemplarily in Fig. 3. The modulator type will just affect the noise part in Eq. (6). Thus for a ΣΔM of order R the output $y(k)$ is determined by

$$y(k) = x(k) + \sum_{r=0}^{R} (-1)^r \cdot \binom{R}{r} \cdot e(k-r) \,. \quad (7)$$

The bandlimited modulator input signal can be represented by a Cardinal series

$$x(k) = \sum_{j=-\infty}^{\infty} x(j \cdot \mathrm{OSR}) \cdot \operatorname{sinc}\left(j - \frac{k}{\mathrm{OSR}}\right) \quad (8)$$

with

$$\operatorname{sinc}(x) = \begin{cases} 1 & ; x = 0 \\ \frac{\sin(\pi x)}{\pi x} & ; x \neq 0 \end{cases} \quad (9)$$

and

$$\mathrm{OSR} = \frac{f_s}{2 f_c} \,. \quad (10)$$

OSR is the oversampling rate. f_s is the sampling rate of the input signal and f_c is its cut-off frequency.

Because of the stationarity of the signal it is sufficient to optimize the filter for one single output value, say $x(0)$. The output $\tilde{x}(0) = \tilde{x}$ of a linear FIR filter can then be calculated by

$$\tilde{x} = \sum_{i=-a}^{a} h(i) y(i) \,, \quad (11)$$

where $h(i)$ is the i^{th} filter coefficient ($i = -a, \ldots, a$). The filter length L is

$$L = 2a + 1 \,. \quad (12)$$

Inserting Eq. (7) into Eq. (11) yields

$$\tilde{x} = \sum_{i=-a}^{a} h(i) x(i) + \sum_{i=-a}^{a} h(i) \sum_{r=0}^{R} \left((-1)^r \cdot \binom{R}{r} \cdot e(i-r) \right) \,. \quad (13)$$

Inserting Eq. (8) for $x(i)$ in Eq. (13) and rearranging terms yields

$$\tilde{x} = \sum_{j=-c}^{c} \left(x(j \cdot \mathrm{OSR}) \cdot \sum_{i=-a}^{a} h(i) \cdot \operatorname{sinc}\left(j - \frac{i}{\mathrm{OSR}}\right) \right) + \sum_{i=-a-R}^{a} e(i) \sum_{r=0}^{R} (-1)^r \cdot \binom{R}{r} \cdot h(i+r) \,. \quad (14)$$

Here we replaced the infinite sum in Eq. (8) by a finite sum. For a high number of elements this is a good approximation. Because of the finite filter length Eq. (14) implies

$$h(i) = 0 \qquad \forall i : \; (i < -a) \vee (i > a) \,. \quad (15)$$

So far the only approximation is that the infinite limits are replaced by the large finite value c. We minimize the MSE

$$\mathrm{MSE} = \mathrm{E}\left\{ (\tilde{x} - x(0))^2 \right\} \,. \quad (16)$$

Inserting Eq. (14) and rearranging results in

$$\mathrm{MSE} = \mathrm{E}\left\{ \left(\sum_{\substack{j=-c \\ j \neq 0}}^{c} x(j \cdot \mathrm{OSR}) \sum_{i=-a}^{a} h(i) \cdot \operatorname{sinc}\left(j - \frac{i}{\mathrm{OSR}}\right) + x(0) \cdot \left(\sum_{i=-a}^{a} h(i) \cdot \operatorname{sinc}\left(\frac{i}{\mathrm{OSR}}\right) - 1 \right) + \sum_{i=-a-R}^{a} e(i) \sum_{r=0}^{R} (-1)^r \cdot \binom{R}{r} \cdot h(i+r) \right)^2 \right\} \,. \quad (17)$$

Now we will additionally assume that $x(j\mathrm{OSR})$ and $e(i)$ ($j = -c, \ldots, c$, $i = -a-R, \ldots, a$) are uncorrelated random variables, i. e. $\mathrm{E}\{e(i')e(i'')\} = 0$, $\mathrm{E}\{x(j' \cdot \mathrm{OSR})x(j'' \cdot \mathrm{OSR})\} = 0$ and $\mathrm{E}\{e(i)x(j \cdot \mathrm{OSR})\} = 0$ ($i' \neq i''$, $j' \neq j''$). We get

$$\mathrm{MSE} = \sigma_s^2 \cdot \sum_{\substack{j=-c \\ j \neq 0}}^{c} \left(\sum_{i=-a}^{a} h(i) \cdot \operatorname{sinc}\left(j - \frac{i}{\mathrm{OSR}}\right) \right)^2 + \sigma_s^2 \cdot \left(\sum_{i=-a}^{a} h(i) \cdot \operatorname{sinc}\left(\frac{i}{\mathrm{OSR}}\right) - 1 \right)^2 + \sigma_e^2 \cdot \sum_{i-a-R}^{a} \left(\sum_{r=0}^{R} (-1)^r \cdot \binom{R}{r} \cdot h(i+r) \right)^2 \quad (18)$$

with

$$\sigma_s^2 = \mathrm{E}\left\{x^2(j \cdot \mathrm{OSR})\right\} \tag{19}$$

$$\sigma_e^2 = \mathrm{E}\left\{e^2(i)\right\} \ . \tag{20}$$

Eq. (18) is to be minimized with respect to $h(i)$ ($i = -a, \ldots, a$). This results in an optimization problem of the type

$$\min_{\mathbf{h}} \|\mathbf{A} \cdot \mathbf{h} - \mathbf{b}\|_2^2 \tag{21}$$

with

$$\mathbf{h} = \left[\ h(-a) \quad \ldots \quad h(0) \quad \ldots \quad h(a) \ \right]^{\mathrm{T}} \tag{22}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \tag{23}$$

and

$$\mathbf{b} = \left[\ 0 \quad \ldots \quad 0 \quad \sigma_s \quad 0 \quad \ldots \quad 0 \ \right]^{\mathrm{T}} \ . \tag{24}$$

The σ_s in vector $\mathbf{b}$ has the index $(c + 1)$. $\mathbf{A}_1$ is a $((2c + 1) \times (2a + 1))$ matrix

$$\mathbf{A}_1 = [A_{m,n}] \tag{25}$$

with

$$A_{m,n} = \sigma_s \cdot \mathrm{sinc}\left((m - c - 1) - \frac{n - a - 1}{\mathrm{OSR}}\right) \tag{26}$$

and $m = 1, \ldots, 2c + 1$, $n = 1, \ldots, 2a + 1$. Matrix $\mathbf{A}_2$ is determined by the order of the modulator. It is a band matrix with the respective binomial coefficients on its diagonals. For instance for the SMI structure it would be

$$\mathbf{A}_2 = \sigma_e \cdot \begin{bmatrix} -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix} \ . \tag{27}$$

σ_s^2 is the parameter of the input signal and can be found by simulations or analysis of the signal $x(k)$. σ_e^2 is the variance of the quantization noise. For the SMI structure and cascades of SMI structures it is

$$\sigma_e^2 = \mathrm{E}\left\{e^2(i)\right\} = \frac{1}{3} \ . \tag{28}$$

Solving Eq. (21) means finding a mean squared solution of the overdetermined linear system of equations

$$\mathbf{A} \cdot \mathbf{h} = \mathbf{b} \ . \tag{29}$$

It is known to be

$$\mathbf{h} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} \mathbf{A}^{\mathrm{T}}\mathbf{b} \ . \tag{30}$$

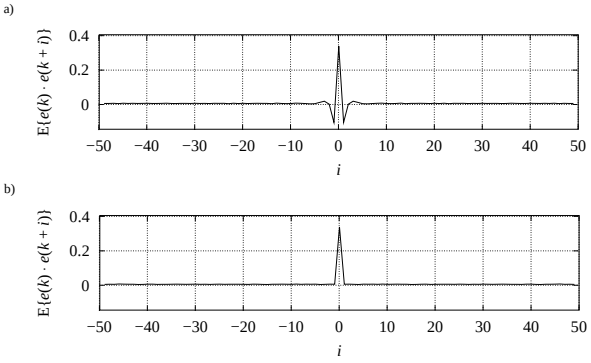

Fig. 4: Autocovariance functions of $e(k)$ for modulators of order a) $R = 1$ b) $R = 2$

Eq. (30) specifies the optimal filter coefficients, i. e. no other linear FIR filter can achieve a higher SQNR. This is only valid if the assumption of uncorrelated random variables is correct.

Calculating the SQNR is straightforward. One merely has to insert the filter coefficients into Eq. (18) in order to obtain the MSE. The signal power is

$$\mathrm{E}\left\{x^2(k)\right\} = \sigma_s^2 \ . \tag{31}$$

This yields

$$\mathrm{SQNR} = \frac{\sigma_s^2}{\mathrm{MSE}} \ . \tag{32}$$

Note again that during the derivation the infinite sum was replaced by a finite sum. This can be easily tolerated, since the sinc-function is converging towards zero. The only essential assumption is that of uncorrelated random variables. If this assumption is correct, then the results in this section are exact. This will be verified in the next section.

3. Results

Here we present results for ΣΔMs of various orders. The first order modulator is the SMI structure in Fig. 2. The higher order modulators were realized by cascading several SMI structures in a MASH structure. Their outputs were connected such that the quantization errors of all stages but the last will cancel out. Fig. 3 shows an example of a second order MASH structure.

First the assumption of uncorrelated random variables was tested by simulations. Fig. 4 shows the autocovariance function of $e(k)$. For the first order modulator there is a correlation of $e(k)$ to its immediate predecessors and successors. For the second order modulator there is no correlation of $e(k)$ to $e(k + i)$ ($i \neq 0$) as can be seen in Fig. 4.b. In fact this is the case for all MASH structures with order larger than one. Further it can be shown that there is no correlation of the quantization error $e(k)$ with the samples $x(j \cdot \mathrm{OSR})$ of the input.

Thus the assumption of uncorrelated random variables is not accurate for the first order modulator. But this is no

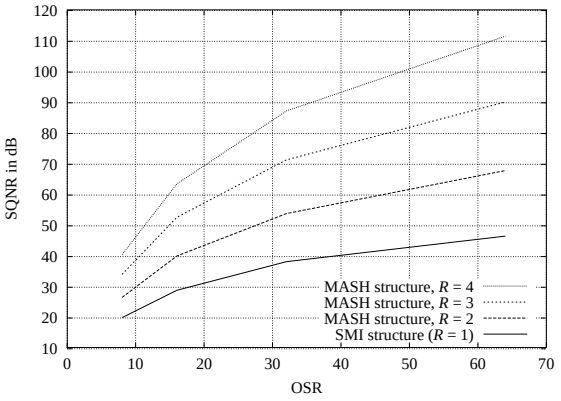

Fig. 5: SQNR vs. OSR for the optimal FIR filter

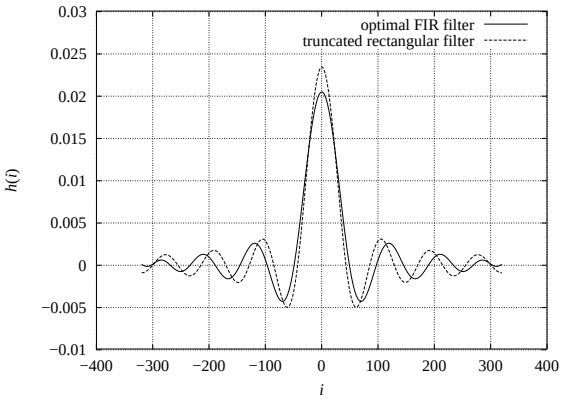

Fig. 6: Filter coefficients of the truncated rectangular and the optimal FIR filter for a MASH structure of order $R = 2$

drawback, since first order modulators are mostly used for simple applications requiring no elaborate filters. However, for MASH structures of order two or higher it is correct. This means the derivations in the previous section are accurate for higher order modulators and the resultant FIR filters are optimal.

Fig. 5 shows the SQNR performance of the filter designed by our method. The filter length was set to $L = 10 \cdot \mathrm{OSR}$. At $\mathrm{OSR} = 64$ our filter gains about 20 dB with each increment of the modulator order.

Designers of ΣΔMs often use filters with a nearly rectangular passband characteristic [5]. In time domain the rectangular filter is a sinc-function. It has infinite length. For practical applications it has to be truncated. The result for a second order MASH structure and $\mathrm{OSR} = 64$ is displayed in Fig. 6 together with the optimal FIR filter. Note that the filters have different zero-crossings. Thus it is not possible to obtain the optimal FIR filter from the truncated rectangular filter by using windowing.

Fig. 7 shows the SQNR performance of the truncated rectangular filter in comparison to the optimal FIR filter designed by our method for a MASH structure of order $R = 2$.

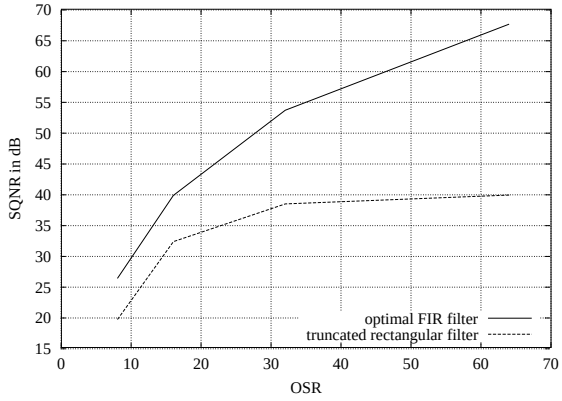

Fig. 7: SQNR vs. OSR for a second order MASH structure

The optimal filter designed by our method performs better than the truncated rectangular filter. At $\mathrm{OSR} = 64$ it gains more than 25 dB.

4. Conclusion

A new method to compute the coefficients of linear FIR filters was presented. The resultant FIR filter is optimal for sigma-delta modulators of order two or higher, i. e. the filter designed by our method achieves the maximum SQNR possible with FIR filters. Thus it is superior to conventional FIR filters. For $\mathrm{OSR} = 64$ it gains about 20 dB with each increment in the modulator order.

Future work will include tests under noisy conditions and the search for convenient realizations.

References

[1] J. C. Candy, Y. C. Ching, and D. S. Alexander, "Using triangularly weighted interpolation to get 13-bit PCM from a sigma-delta modulator," *IEEE Transactions on Communications*, vol. COM-24, no. 11, pp. 1268–1275, November 1976.

[2] R. M. Gray, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input," *IEEE Transactions on Communications*, vol. 37, no. 6, pp. 588–599, June 1989.

[3] I. Wiemer, "Linear and Nonlinear Decoding Algorithms for the First Order Sigma-Delta Modulator," diploma thesis, Dresden University of Technology (German), December 2002.

[4] L. L. Presti, "Efficient modified-sinc filters for sigma-delta A/D converters," *IEEE Transactions on Circuits and Systems-II*, vol. 47, no. 11, pp. 1204–1213, November 2000.

[5] S. R. Norsworthy, R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters: Theory, Design and Simulation*. IEEE Press, 1997.