means that the transition from stable to unstable operation is 'sharp', i.e., there is no extended amplitude range where the modulator is only weakly unstable with a very low escape rate. A low escape rate means that the mean survival time is long but finite, i.e., the modulator is seemingly stable for most short simulations. It is obvious that a modulator intended for a general purpose data converter cannot be tested for all possible input signals by simulations. However, if the modulator is absolutely stable for every constant input in the actual input amplitude range, there is a good chance that the modulator also can handle time varying signals with a suitable band-limitation. Even an exhaustive simulation test with constant inputs requires a lot of computational power.

The optimization methods presented become actually more and more necessary as the dimension increases further because the NTF-prototype method results in more and more unreliable modulators.

7.2 The Choice of Parameter Space

The coordinates for the $(N-1)$-dimensional parameter space of an Nth order modulator with fixed feedback filter poles can be chosen in many ways. The actual z-domain polynomial coefficients are a bad choice due to the extreme sensitivity to the zero locations. A better approach is to describe the zero locations directly. This can be accomplished by a suitable transform. E.g., for a fourth order modulator, the three zero locations z_1, z_2 and z_3 can be found using the transform:

$$z_1 = t_1 + \sqrt{t_2}\,, \quad z_2 = t_1 - \sqrt{t_2}\,, \quad z_3 = t_3 \tag{7.1}$$

where t_1, t_2 and t_3 are real parameters.

The third zero z_3 is always real and equal to t_3. The zero pair z_1 and z_2 are complex conjugates for negative t_2-values and real valued for positive t_2. There is a one-to-one correspondence between the three real valued parameters and the three zero locations of which two might be complex. Furthermore, there is a 'smooth' transition from complex to real zero locations.

The parameter to zero location transform of Eq. (7.1) can easily be extended to any dimension.

7.3 BIBO stability

The Gaussian stability criterion presented in Sec. 6.5 was related to the noise power amplification factor $\mathrm{A}(K)$ which is the squared two-norm of the noise transfer function of the linearized system. This stability criterion was derived by assuming the signals in the modulator to be stochastic signals. Hence, the Gaussian criterion is based on mean (squared) signal values. As stated previously, such criterion is only approximate: it is not sufficient and nor necessary. The purpose of this Section is to discuss a *bounded input bounded output — BIBO* criterion which is based on a worst case assumption.

For conceptual reasons it is convenient to transform the usual Σ-Δ modulator to a new equivalent circuit which is shown in Fig. 7.1. The circuit consists of an input prefilter $F(z)$ modifying the input to the signal $\hat{x}(k)$ which is fed to a so-called Noise-Shaper [38, 41, 62, 63]. The quantization noise $q(k)$ is found by subtracting the quantizer output from the quantizer input $e(k)$ and the quantization noise is fed back through a filter $F(z)$

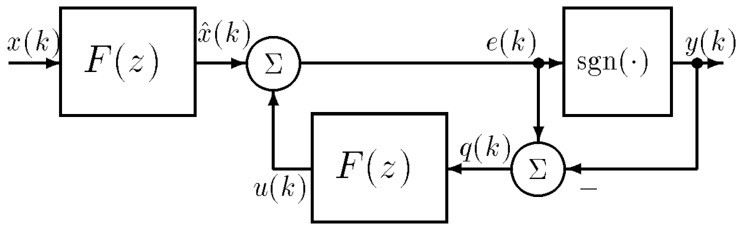

Figure 7.1: A Noise Shaper with a linear prefilter. The system is equivalent to a usual Σ-Δ modulator.

and added to $\hat{x}(k)$ in order to form $e(k)$. A circuit analysis reveals that the entire system is equivalent to a usual Σ-Δ modulator having a feedback filter:

$$H(z) = \frac{F(z)}{1 - F(z)} \tag{7.2}$$

Having $H(z)$ as starting point, the $F(z)$ filter can be found as:

$$F(z) = \frac{H(z)}{1 + H(z)} \tag{7.3}$$

Since $H(z)$ usually has a very high low frequency gain, it is seen that $F(z = 1) \approx 1$. This means that $F(z)$, as prefilter regarded, is quite harmless at low frequencies.

The first step for the BIBO criterion is to assume that the maximum magnitude of the quantization noise is below unity, i.e., $\max(|q(k)|) = \|q(k)\|_\infty \leq 1$. This assumption is only justified as long as the quantizer input magnitude $e(k)$ is below 2. Using a worst case limit of the quantizer input magnitude the BIBO criterion becomes:

$$\|f(k)\|_1 \cdot \|q(k)\|_\infty + \|\hat{x}(k)\|_\infty \leq 2 \tag{7.4}$$

where $\|f(k)\|_1$ is the one-norm of the impulse response of the $F(z)$ filter, i.e., the sum of the magnitude of the impulse response:

$$\|f(k)\|_1 = \sum_k |f(k)| \tag{7.5}$$

When Eq. (7.4) is fulfilled, the maximum quantization noise magnitude is unity:

$$\|\hat{x}(k)\|_\infty \leq 2 - \|f(k)\|_1 \tag{7.6}$$

This gives an upper bound on the stable input amplitude range, i.e., the one-norm of $F(z)$ must be as low as possible. The criterion of Eq. (7.6) is sufficient but rather conservative in most cases [58, 59]. This one-norm based stability criterion was proposed in [4] and further improved in [60].

The BIBO criterion can be improved (i.e., made less conservative) by using the modulator invariance to positive scalings of the feedback filter $H(z)$ [59]. Hence, $H(z)$ can be scaled by a ny positive number K in Eq. (7.3) in order to get different $F(z)$ filters resulting in exactly the same modulator with exactly the same stability properties:

$$F_K(z) = \frac{K \cdot H(z)}{1 + K \cdot H(z)} \tag{7.7}$$

Comparing to Eq. (6.3) it is seen that:

$$F_K(z) = \frac{1 + K \cdot H(z)}{1 + K \cdot H(z)} - \frac{1}{1 + K \cdot H(z)} = 1 - \mathrm{NTF}_K(z) \tag{7.8}$$

Notice that the filter $F_K(z)$ is equal to the signal transfer function $\mathrm{STF}_K(z)$, cf. Eq. (6.2) for quantizer gain K and $G(z) = H(z)$; however in this context K is just an arbitrary scaling factor.

The one-norm of $\mathrm{NTF}_K(z)$ is defined as $S(K)$:

$$S(K) \triangleq \|\mathrm{ntf}_K\|_1 = \sum_k |\mathrm{nft}_K(k)| \tag{7.9}$$

Since $\mathrm{ntf}_K(0) = 1$, it is noted that:

$$\|f_K(k)\|_1 = S(K) - 1 \tag{7.10}$$

This relationship leads to an improved BIBO criterion:

$$\|\hat{x}(k)\|_\infty \leq 3 - \mathrm{S}_{min} \tag{7.11}$$

where S_{min} is the global minimum of $S(K)$.

The improved BIBO criterion is quite analogous to the Gaussian criterion: the BIBO criterion is based on the minimum one-norm of the NTF and the Gaussian criterion is based on the minimum two-norm of the NTF. A guaranteed zero-input stable modulator has $\mathrm{S}_{min} \leq 3$ and corollary: $\mathrm{A}_{min} \leq 3^2$. Since the leading ntf_K term is unity, it is concluded that $\mathrm{A}_{min} \leq (1 + 2^2)$. The Gaussian criterion suggests that $\mathrm{A}_{min} < 2.75$ for a zero input stable modulator. The difference between the norms is that the two-norm gives more weight to the big elements of the impulse response due to the squaring. The one-norm gives equal weights to all elements. This difference is reflected in the two criteria: the Gaussian two-norm based criterion works fine in many cases but it gives no guarantee what soever. The one-norm based BIBO criterion is ironclad but often too conservative because it is based on the worst case extremal quantization noise sequence. It is often very unlikely that this worst case sequence will occur in practise. The BIBO criterion can be further improved by including exact knowledge of the dependence between adjacent quantizer input samples. However, this approach is difficult and not very general. Such methods try to identify a trapping region in the state-space, i.e., a orbits starting inside the trapping region never leave it. Specialized stability tests based on these principles have been developed for second order modulators [67, 59]; see also Example 3.4.

Example 7.1 This example demonstrates the BIBO criterion for some simple first- and second-order modulators: the chaotic first order modulator with $H_1(z) = z^{-1}/(1 - 2z^{-1})$ and the usual second order modulator with $H_2(z) = (z^{-1} - 0.5z^{-2})/(1 - 2z^{-1} + z^{-2})$. The $S(K)$-curves for these two filters are shown

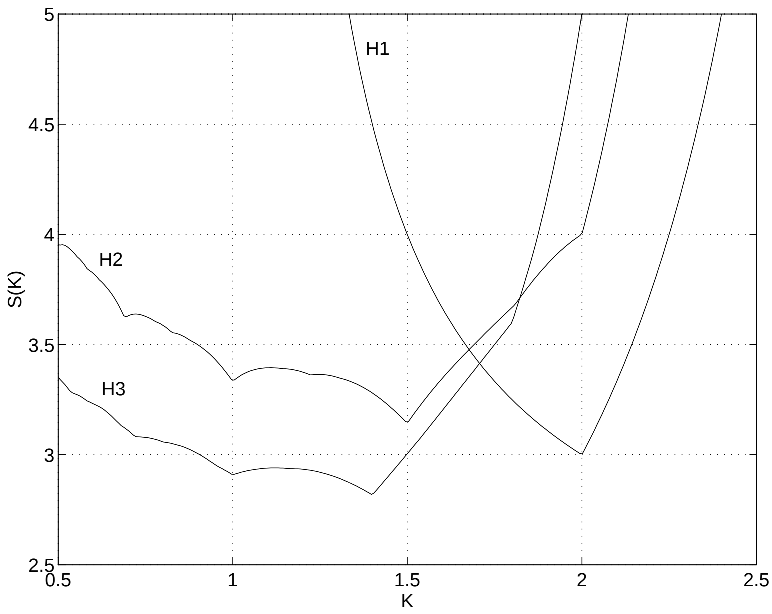

Figure 7.2: $S(K)$-curves for three feedback filters: $H_1(z) = z^{-1}/(1-2z^{-1})$, $H_2(z) = (z^{-1}-0.5z^{-2})/(1-2z^{-1}+z^{-2})$ and $H_3(z) = (z^{-1}-0.6z^{-2})/(1-2z^{-1}+z^{-2})$

on Fig. 7.2. The first order modulator $H_1(z)$ is marginally stable since $S(2) = \mathrm{S}_{min} = 3$. In this special case the Gaussian criterion is more conservative that the BIBO criterion.

The BIBO criterion cannot guarantee the stability of the popular second-order modulator $H_2(z)$ since $\mathrm{S}_{min} > 3$. Both $H_1(z)$ and $H_2(z)$ have noise transfer functions that become FIR filters for a suitable K-value, i.e., $K = 2$ in this case. There is a general class of feedback filters having this property. Every second-order modulator with this FIR-property conforms to Eq. (3.5). Every first order modulator has the FIR property, and furthermore, the S_{min}-value is taken when the noise transfer function is FIR. The second-order modulator $H_2(z)$ has $S(2) = 4 > \mathrm{S}_{min}$ in the FIR-case, i.e., the scaling giving FIR NTF is not the optimal scaling.

The zero of $H_2(z)$ can be modified slightly such that the second-order modulator is guaranteed stable for low input amplitudes, e.g., the $S(K)$ curve of $H_3(z) = (z^{-1}-0.6z^{-2})/(1-2z^{-1}+z^{-2})$ is also plotted on Fig. 7.2. This modulator has $\mathrm{S}_{min} \approx 2.82$ giving a guaranteed stable maximum input amplitude of 0.18. This stability enhancement takes place at the expense of a deteriorated in-band noise suppression due to a deteriorated low-frequency loop-gain □

7.4 Reliability

A practical constraint in feedback filter optimization is that the modulator is suitably reliable, i.e., the onset of instability occurs abruptly when the input amplitude is increased. The reliability ensures that the MSA values found by simulations using slowly

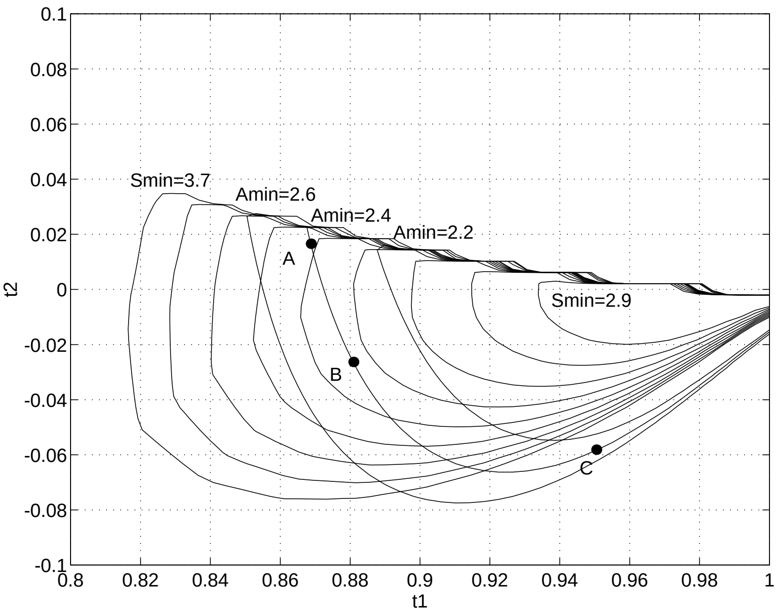

Figure 7.3: Contours of equal S_{min} and A_{min} for $t_3 = 0.7$. The parameters t_1, t_2 and t_3 determine the feedback filter zeros, cf. Eq. (7.1). The pole locations are given by Eq. (7.12). The S_{min} contours are shown for increments of 0.1 from 2.9 to 3.7. The A_{min}-contours are for 2.2, 2.4 and 2.6. Three points (A,B and C) are marked on the $A_{min} = 2.4$ contour.

increasing ramp input will hold in practise when the modulator operates at full speed (e.g., 3 MHz sample rate) with input signals from 'the real world'. A reliable modulator gives a simulated MSA which is fairly independent of the dc sweep rate. Conversely, an unreliable modulator can exhibit an acceptable MSA for high sweep rates corresponding to short simulations — but longer simulations may reveal instability for much lower input amplitudes.

The question is now: how can feedback filters of reliable and unreliable modulators be distinguished from each other? Numerous simulations have indicated that the two-norm based Gaussian criterion predicts the onset of severe instability fairly well, i.e., the Gaussian criterion is consistent for high sweep rates. However, different filters with the same A_{min}-value may behave differently with respect to reliability.

The existence of extended input ranges with low escape rates is one of the characteristics of unreliable modulators. A low escape rate means that the system only escapes when the orbit enters certain rarely visited regions in the state-space, i.e., it is a worst case situation which occurs rarely. This situation might arise when the minimum one norm S_{min} is composed of many small contributions, i.e., when the NTF impulse response has slowly decaying (high-Q) oscillations. The worst case quantizer input can be very large but it is, on the other hand, very seldom that values close to the extremal occur. A more reliable modulator is expected to exhibit a suitably low S_{min}-value dominated by the first few terms of the NTF impulse response.

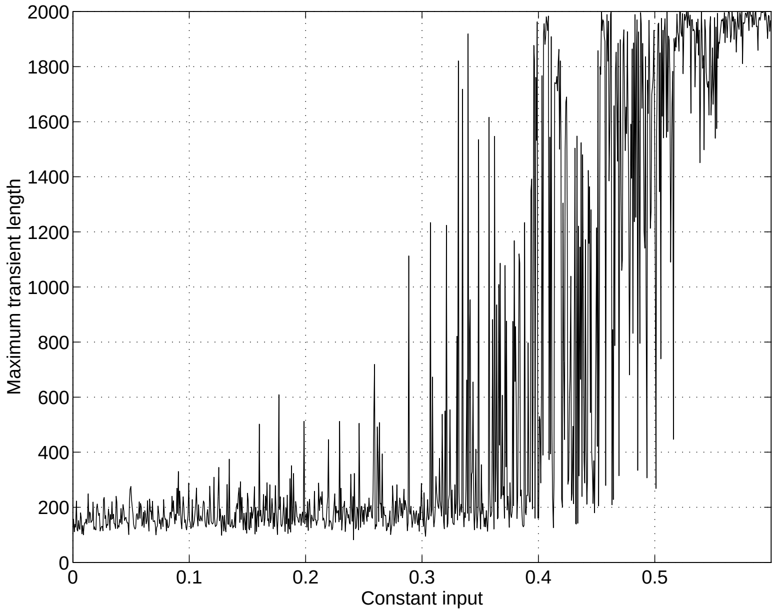

Figure 7.4: Plot of maximum transient length versus constant input for the modulator A, cf. Fig. 7.3. For simulation parameters see Example 7.2.

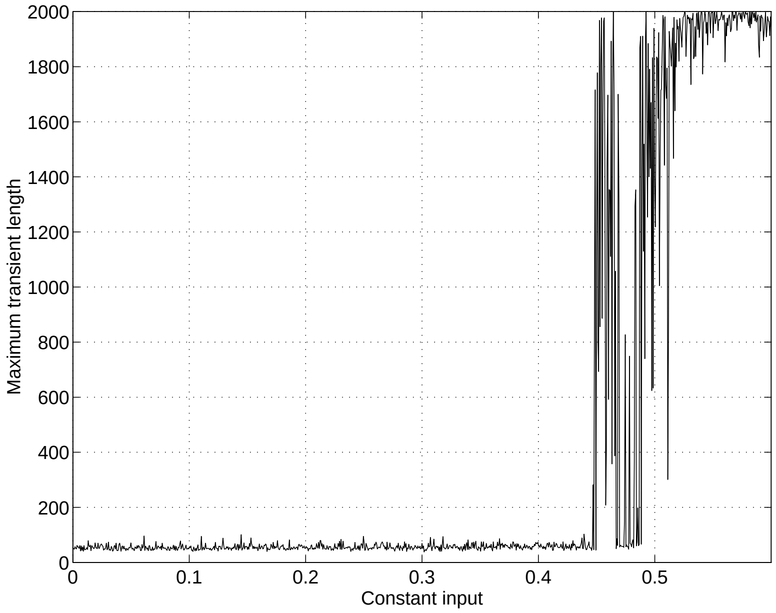

Figure 7.5: Plot of maximum transient length versus constant input for the modulator B, cf. Fig. 7.3. For simulation parameters see Example 7.2.

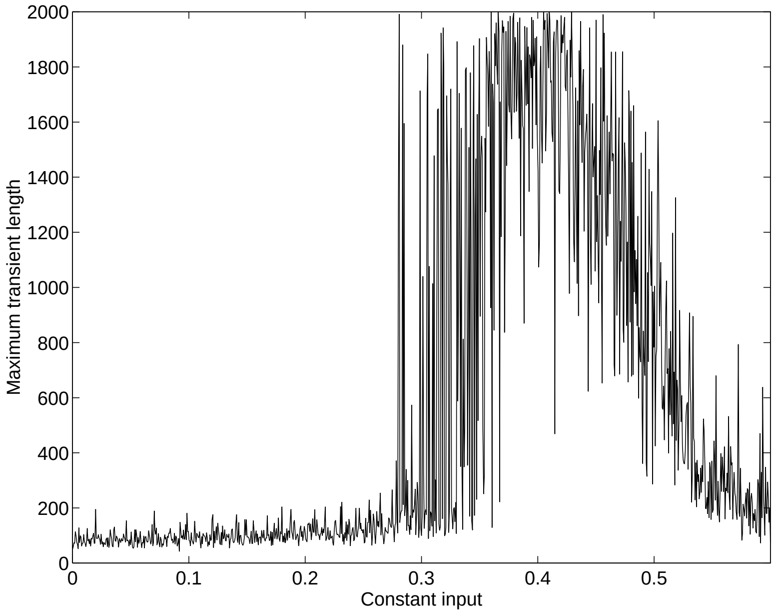

Figure 7.6: Plot of maximum transient length versus constant input for the modulator C, cf. Fig. 7.3. For simulation parameters see Example 7.2

Example 7.2 This example demonstrates the behavior of three fourth order modulators with $\mathrm{A}_{min} = 2.4$. The feedback filter poles are suited for 64 times oversampling, i.e., base-band $f_b = 20\,\mathrm{kHz}$ and sample rate $f_s = 64{\cdot}48\,\mathrm{kHz}$. The four optimized unit circle feedback filter poles are located at the frequencies [58]:

$$f_i \in \left\{ \pm\sqrt{\frac{3}{7} \pm \sqrt{\left(\frac{3}{7}\right)^2 - \frac{3}{35}}} \cdot f_b \right\} \tag{7.12}$$

These pole frequencies should minimize the unweighted total base-band noise power.

The (t_1, t_2, t_3) parameter space was investigated with the optimized poles for $f_b = 1/64 \cdot 20/24$ relative to half the sample rate. Fig. 7.3 shows two-dimensional contours for equal S_{min} with t_3 fixed to 0.7. The graph includes the contour for $\mathrm{A}_{min} = 2.2$, 2.4 and 2.6. There is a lot of similarity between the types of contours: the gradients are in almost the same directions, i.e., both the A_{min} and S_{min} measures will in most cases agree on whether a given parameter change enhances the stability or not. However, the two stability measures are still quite different in nature. This will be examined by following one of the A_{min}-contours. Three points A, B and C on the 2.4 contour are marked. Point B has minimum $\mathrm{S}_{min} = 3.24$ for $t_3 = 0.7$ and point A and C have somewhat higher values, namely 3.31 and 4.7, respectively. The modulators of the three points are not guaranteed stable for zero input according to the BIBO criterion but they should all have an MSA of approx. 0.38 according to the Gaussian criterion of Eq. (6.20).

The stability of the modulators was investigated in two ways: simulations with slowly increasing input ramp and an exhaustive search for long transients. The latter method was introduced in Sec. 5.2. The modulator to be tested is simulated for a large number of random initial conditions and randomly selected constant inputs for a maximum number of time-steps. If the quantizer input magnitude exceeds an upper bound of 10, the current number of time steps simulated is regarded as a transient length. The maximum transient length is found within each of a number of constant input bins. A plot of the maximum transient length versus the constant input magnitude gives a kind of fingerprint of the stability and reliability of the modulator. Such plots are shown on Fig. 7.4, 7.5 and 7.6 for the three modulators corresponding to point A, B and C. Each of the plots are obtained from 5 mill. simulations with random constant input distributed uniformly over the interval $[0, 0.6]$. The initial conditions were randomly selected using a uniform distributions in the quantizer input delay-space (i.e., the coordinates $\boldsymbol{e}_0 = [e(-1), ...e(-4)]$) such that $\|\boldsymbol{e}_0\|_\infty = 0.5$. The simulations were stopped either when the quantizer input exceeded a magnitude of 10 or after a maximum of 2000 time-steps. The maximum transient length was found within 1000 constant input bins.

The three maximum transient plots convey a lot of information: modulator B is very nice and is expected to operate reliably for input magnitudes up to approx. 0.45 where there is a transition region between 0.45 and 0.5. For magnitudes above 0.5 the modulator is definitely unstable. The maximum transient 'background level' is low and very constant for magnitudes less than 0.45. For this rather reliable modulator the Gaussian criterion is somewhat pessimistic.

The plots for modulator A and C have a quite different behavior: the graphs have a lot of spikes of different sizes for most of the entire input range. Furthermore, the background levels for both graphs are significantly higher than for modulator B. It is obviously difficult to judge the MSA for these very unreliable modulators. Simulations with ramp type input rising from zero to unity during 10^8 samples (corresponding to approx. 30 seconds real-time) have revealed that modulator B becomes unstable at input magnitudes as low as 0.33 and 0.26 for modulator C. Modulator B can handle input magnitudes up to 0.44 even for these very long ramp simulations. Such slowly rising ramp input is realistic for signals from, e.g., temperature and pressure transducers. The question is of course how these three modulators will behave generally in a real application. There is only one way to give an exhaustive answer: implement the circuits and test the devices for extremely long time. However, it seems likely that the more reliable modulator B is absolutely stable when operating with suitably band-limited signals having a maximum magnitude below 0.4.

Another observation from the maximum transient simulations was that the fraction of initial conditions surviving the maximum 2000 time steps vary from modulator to modulator. Fig. 7.7 shows the surviving fraction within each of the 1000 input bins for modulator B and C. It is seen that the surviving fraction for low input magnitude is almost 5 times higher for modulator B than for modulator C and this indicates that the volume of the basin of attraction for the unreliable modulator C is lower, i.e., it is more difficult to find stable initial conditions.

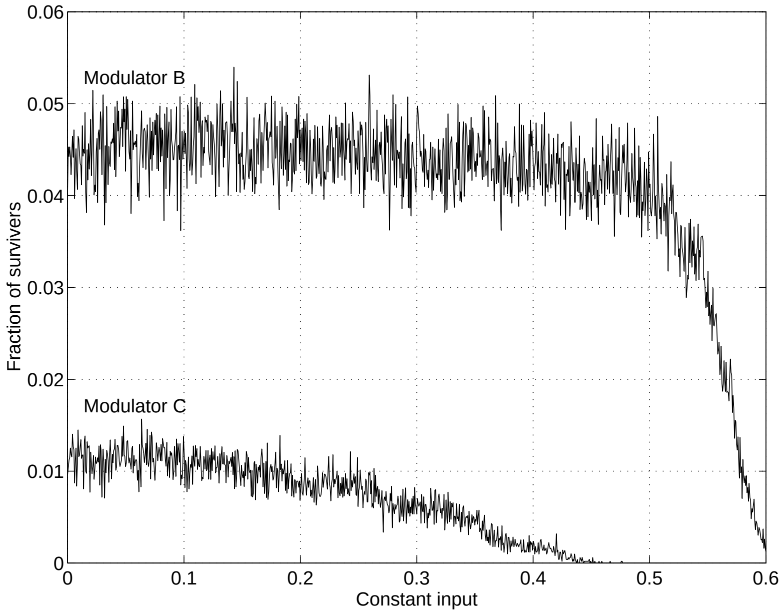

Figure 7.7: Fraction of initial conditions surviving 2000 time steps for the modulators B and C, cf. Fig. 7.3. For simulation parameters see Example 7.2

The surviving fraction for modulator B is fairly constant up magnitudes near 0.5. For higher magnitudes the fraction decreases fast to near zero for magnitudes around 0.6. This reflects a sharply increasing escape rate in this transition region. □

The previous example demonstrated the behavior of reliable and unreliable modulators. The very high S_{min}-value for modulator C can clearly be taken as an indication of unreliability. In fact, the most reliable modulator had the lowest S_{min}-value. More detailed information is probably conveyed by the entire NTF impulse response which corresponds to the S_{min}-value. In fact, the NTF impulse response fully describes the modulator and the question of stability is therefore 'just' a matter of proper interpretation.

Example 7.3 The $\mathrm{ntf}_{min}(k)$ impulse responses that gives the S_{min}-values for the modulators A, B and C of Fig. 7.3 are plotted on Fig. 7.8. By definition $\mathrm{ntf}_{min}(0) = 1$ due to the one sample feedback delay. Furthermore, $\mathrm{ntf}_{min}(1) = -Kc_1$ where c_1 is the first $H(z)$ numerator term which is normalized to unity in this example. Consequently, for the shown plots, $\mathrm{ntf}_{min}(1) = -K_{Smin}$ where K_{Smin} is the K-value that yields S_{min}. The K_{Smin}-values are all near 1.5 for the three modulators. Hence, the first two dominating terms make up a sum of approx. 2.5 and this shows that it is the 'tails' of the NTF impulse responses which are responsible for the instability and possible unreliability. First order modulators have FIR NTF, i.e., they have no 'tails' and the onset of instability is consequently very abrupt.

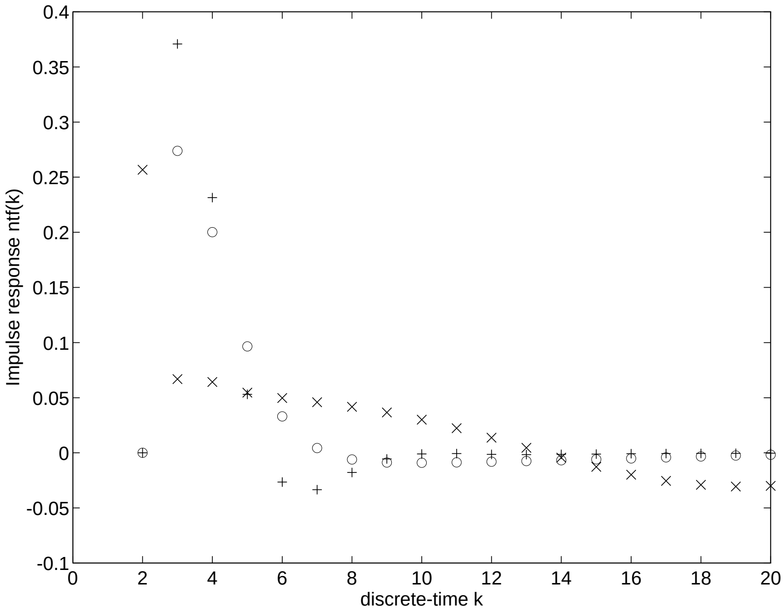

Figure 7.8: Plot of the $\mathrm{ntf}_K(k)$ impulse responses which achieves the S_{min}-values for modulator A(+), B(o) and C($\times$), cf.Fig. 7.3. The plot shows only the tails for $k \geq 2$.

For both modulator A and B, the third term (i.e., $\mathrm{ntf}_{min}(2)$) is zero and non-zero for modulator C. This is probably a consequence of the one-norm minimization.

The very high S_{min}-value for modulator C is due to a slowly decaying oscillation which probably causes the unreliability. It is more difficult to explain the cause of unreliability for modulator A. However, a comparison between the NTF impulse responses of modulator A and B reveals that the 'tail' of modulator B decays to zero without oscillation while the 'tail' of modulator A is more oscillating before it dies out. The worst case occurs when the past quantization noise values have signs which fit the signs of the NTF impulse response. This can cause the quantization noise to increase and blow up the modulator. Modulator B is probably more reliable and stable because of the almost unipolar tail, i.e., it is very unlikely, in normal stable operation, that a modulator produces long sequences of unipolar output codes. Conversely, oscillating code patterns are more likely to occur and this can produce a worst case situation for modulators with oscillating tails $\square$

The simple one-norm S_{min} can explain some but not all of the differences in reliability. The observations done in the previous example suggests that the NTF impulse responses should be weighted suitably in order to provide a measure which indicates the degree of reliability. A possibility is to examine the NTF impulse response in the spectral domain. A necessary condition for BIBO stability is that the maximum magnitude of the filter $F_K(z)$ is below 2, i.e., when the one-norm of $f_K(k)$ is below 2 as required for BIBO stability, the magnitude response of $F_K(z)$ is likewise bounded by 2.

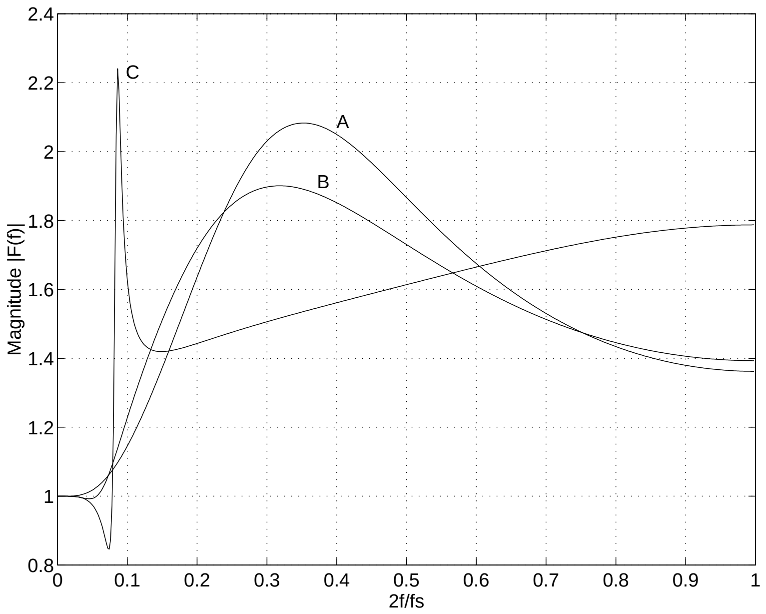

Figure 7.9: Magnitude responses of the $F_{min}(z)$ filters with minimum one-norm for modulator A, B and C, cf. Fig. 7.3.

Example 7.4 On Fig. 7.9 the magnitude responses of the $F_{min}(z)$ filters with minimum one-norm for the A,B and C modulators (se Fig. 7.3 are plotted. It is seen that modulator B is the only modulator with a magnitude maximum below 2. The oscillating behavior of modulator C is seen as a strong peak with a magnitude higher than 2. The oscillating behavior is also seen on Fig. 7.10 which shows the simulated output spectrum of modulator C. This spectrum has clearly a similar strong peak. The maximum magnitude for modulator B is well below 2 and this probably explains the better reliability. □

The previous example showed that the maximum magnitude of $F_K(z)$ is a good measure of reliability. This motivates the definition:

$$\mathrm{F}_{max} = \max\left\{\left|F_{min}(e^{j\pi f})\right|\right\} = \|F_{min}(z)\|_\infty \tag{7.13}$$

The F_{max} measure is thus the infinity norm of the $F_{min}(z)$ filter having minimum one-norm.

Example 7.5 Fig. 7.11 shows contours of constant F_{max} for $t_3 = 0.7$. The remaining parameters conform to Example 7.2 and Fig. 7.3 □

The three measures A_{min}, S_{min}, and F_{max} are based on the two, one and infinity norms, respectively. The combination of these three figures gives a quite good picture of the stability of a modulator. The unreliable modulators are typically characterized by the F_{max} and S_{min} values being too high compared to the A_{min} value.

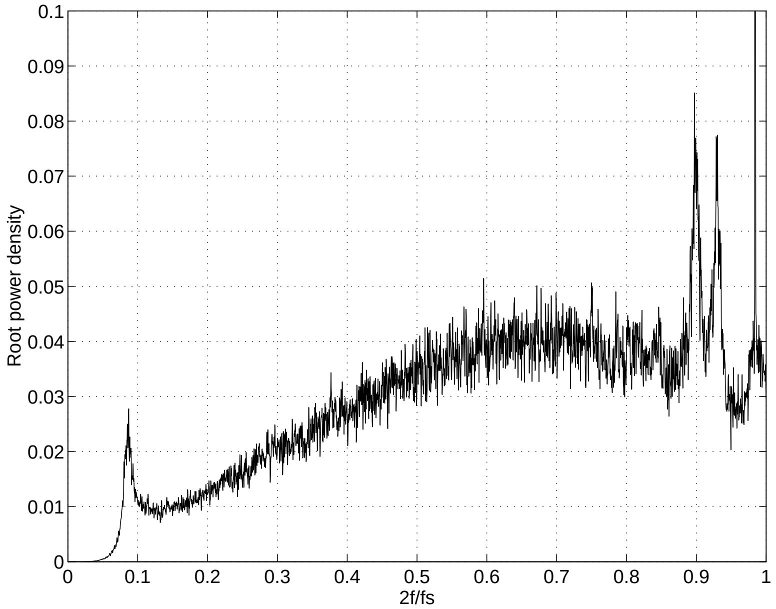

Figure 7.10: Power spectrum estimate for the simulated output of modulator C, cf. Fig. 7.3. A constant input of 1/64 was used. Notice the linear axes and the peak near $2f/f_s \approx 0.095$

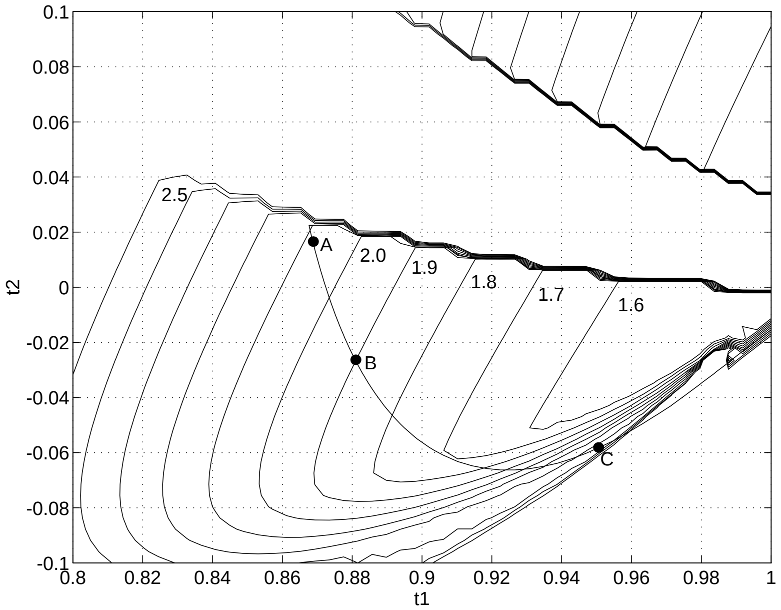

Figure 7.11: Contours of constant F_{max} for $t_3 = 0.7$. The $A_{min} = 2.4$ contour is plotted and the points A,B and C, cf. Fig. 7.3, are marked.

7.5 Performance Prediction

Optimization of loop filters will normally comprise a minimization of the in-band quantization noise power. A brute force method is to simulate the system and estimate the in-band noise power using a filter or an FFT spectral analysis. However, this approach can be too time consuming for use in connection with an automated optimization. Furthermore, for a finite length simulation, the obtained in-band noise power as a function of the filterparameters is not very 'smooth' due to the stochastic nature of the modulator. This effect can literally 'derail' an optimization routine which is based on numerically estimated gradients.

The use of linearized modeling can provide a computational shortcut to performance prediction. The linearized loop model discussed in Ch. 6 shows how the quantization noise is shaped and suppressed in the base-band. The apparent problem is that either the quantizer gain K or the feedback filter scaling can be chosen arbitrarily in such models. It has been suggested to use a fixed quantizer gain of 2 for a feedback filter scaling where the leading term of the feedback filter impulse response is unity [56].

The next problem is to predict the variance and spectrum of the quantization noise in order to compute the base-band noise power. The quasilinear modeling framework offers a fairly simple and yet very precise way to predict the noise spectrum of the modulator: once the equilibrium point (e.g., the quantizer gain K) is known, the quantization noise power as well as the noise transfer function is known and the in-band noise power can subsequently be found by integration. It was demonstrated in Sec. 6.8 that the equilibrium point can be predicted very well using an algorithm which predicts the quantizer input probability distribution function. However, this algorithm is far to complex and slow for use in a loop filter optimization procedure.

A simpler approach is to use the fact that most modulators approximately follows the same $\mathrm{A}(m_y)$-curve: for zero input, most modulators operate with equilibrium $\mathrm{A}_{eq} \approx 3.49$ (see Fig. 6.10). This rule of thumb applies to most 'good' and 'well behaved' modulators. The rule is typically less accurate for modulators having a very high stable amplitude range.

For a K_{eq} satisfying $\mathrm{A}(K_{eq}) = \mathrm{A}_{eq}$ the base-band quantization noise power σ_b^2 for zero input is given by:

$$\sigma_b^2 = \frac{1}{\mathrm{A}_{eq}} \int_0^{f_b} \left| \frac{1}{1 + K_{eq} \cdot H(e^{j\pi f})} \right|^2 df \tag{7.14}$$

where f_b is the upper base-band limit frequency relative to half the sample rate.

This gives an estimate of the in-band noise power for small amplitude input. The in-band noise power will increase somewhat for higher amplitudes.

Example 7.6 Fig. 7.12 shows contours of equal σ_b^2, cf. Eq. (7.14) with $\mathrm{A}_{eq} = 3.49$, in the (t_1, t_2) parameter plane with $t_3 = 0.7$. The feedback filter poles are given by Eq. (7.12) with $f_b = 1/64 \cdot 20/24$. The points A, B and C, cf. Fig. 7.3, are marked. The three modulators were simulated with a constant input of 1/64 and the output spectra was estimated using averaged and windowed 8k FFT's. The small dc-bias ensures that possible tones are outside the base-band. The base-band noise power was estimated by summing up the power of

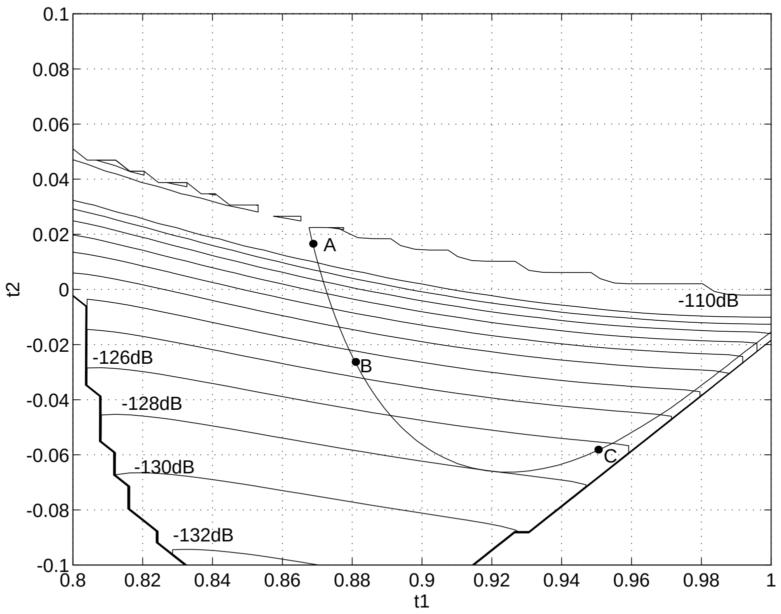

Figure 7.12: Contours of predicted σ_b^2 for $t_3 = 0.7$. The $\mathrm{A}_{min} = 2.4$ contour is also plotted and the points A,B and C, cf. Fig. 7.3, are marked.

the first $f_b \cdot 4096$ FFT-bins. The simulated and predicted values of σ_b^2 appears from the table:

Modulator	Predicted σ_b^2	Simulated σ_b^2
A	-96.8 dB	-103.8 dB
B	-122.9 dB	-124.8 dB
B	-126.4 dB	-127.7 dB

The predicted σ_b^2 is clearly to high in all cases. Especially for modulator A, the prediction is 7 dB to pessimistic. However, for modulators with reasonably high performance, the prediction is consistently only a few dB wrong.

It is seen that point A gives the highest in-band noise power and this is due to the feedback filter zeros being all real valued. The highest zero modulus is 0.998 and this almost cancels the effect of the dc-pole at $z = 1$. Furthermore, this modulator is also proven to be very unreliable. Point B and C with negative t_2 have a complex zero pair giving a higher low frequency loop gain and thereby a better in-band noise rejection. The difference between B and C is only a few dB. When the known unreliability and lower MSA of point C is taken into consideration, the modulator of point B must clearly be preferred.

Following the $\mathrm{A} = 2.4$ contour on Fig. 7.12, is can be seen that a σ_b^2 minimum is found near the point $(0.92, -0.066)$. It appears from Fig. 7.3 and Fig. 7.11 that this point also gives acceptable stability and reliability properties $\square$

Another useful approach for prediction of the equilibrium point, is that the K-value which minimizes the one-norm $\mathrm{S}(K)$ is often quite close to the equilibrium K found by simulations.

Another interesting observation is that the two-norm of the error transfer function $\mathrm{ETF}_K(z)$ (see Eq. (6.26)) has minimum close to the equilibrium K. Recall that the ETF given by Eq. (6.26) is the transfer function from the quantization noise source to the quantizer input, i.e., the modulator attempts to operate with minimum quantizer input variance. From the definition of the noise amplification factor A of Eq. (6.7) it can be derived that:

$$\|\mathrm{ETF}_K\|_2^2 = \frac{\mathrm{A}(K) - 1}{K^2} \tag{7.15}$$

The K-value which minimizes Eq. (7.15) can thus be used as K_{eq}. In some cases this approach gives very good performance predictions.

Example 7.7 The fifth order Chebychev II derived filter A from Table 6.2 has a very low zero input A-value. The minimum two-norm of the ETF_K is reached for $K = 2.05$ and this K gives A= 2.74. A simulation with zero input yields $K = 2.09$ and A= 2.86. However, Fig. 6.10 shows that for increasing input amplitude, the noise amplification factor is increasing towards a maximum above 3.5 for $m_y \approx 0.12$ □

7.6 Optimization Strategies

A loop filter optimization procedure should minimize the in-band noise power under suitable constraints. The constraints should first of all ensure that the modulator is stable and reliable in a prescribed amplitude range. A widely used constraint has been to impose an upper bound for the unity quantizer gain noise amplification factor $\mathrm{A}(K = 1)$ [3]. Other approaches have been to impose an upper bound on the maximum amplitude gain of the noise transfer function as used for NTF-prototypes in [10].

A better approach is to specify an upper bound for A_{min} determined by the Gaussian criterion and the desired MSA. This constraint takes the feedback filter scaling invariance into consideration. An optimization with this constraint alone will in the general case lead to an unreliable modulator with an MSA lower than prescribed by the Gaussian criterion. Consequently, suitable upper bounds for S_{min} and F_{max} must normally be specified in order to get a good result. In many cases, only some of the constraints are active at the same time, i.e., restricting the optimization. The rest of this section will demonstrate how a suitable mixture of one-, two- and infinity-norm based constraints leads to consistent results.

An optimization procedure employing the mentioned constraints was implemented in MATLAB using the Optimization Toolbox CONSTR routine [21] based on Sequential Quadratic Programming [1]. The approach is fairly robust and has been applied successfully to modulators with orders ranging from 3 to 8. A similar approach called CLANS (Closed Loop Analysis of Noise-Shaping coders) using CONSTR has been reported in [29] where a suitably weighted in-band noise power measure is minimized under a one-norm constraint. This approach was successfully used for optimized design of multibit noise-shapers. However, CLANS is not intended for one-bit modulators and the approach does not take the scaling invariance into consideration.

[1]The MATLAB tools are available from the author

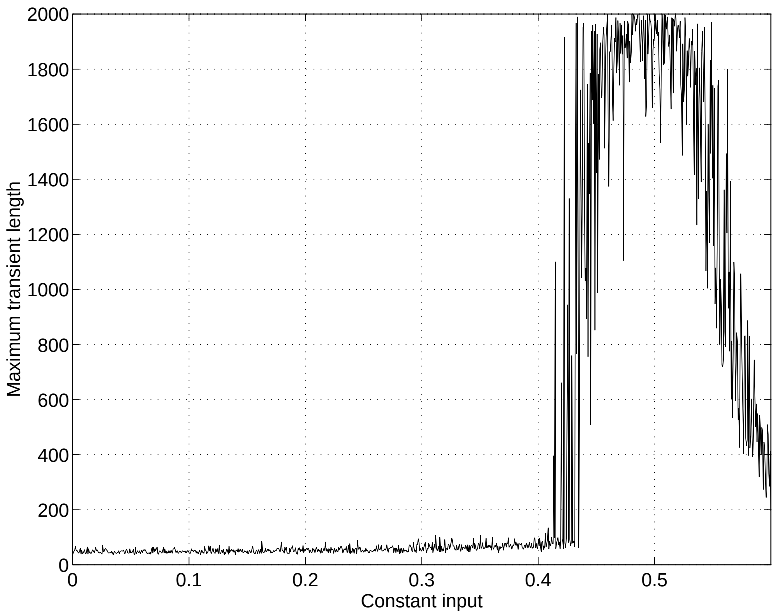

Figure 7.13: Maximum transient plot of an optimized modulator with the constraints $A_{min} < 2.4$, $F_{max} < 1.8$ and $S_{min} < 3.5$.

Example 7.8 A fourth order modulator was optimized for $f_b = 1/64 \cdot 20/24$. The feedback filter poles was found from Eq. (7.12) and the zeros were obtained from a constrained minimization of the predicted σ_b^2 using the constraints $A_{min} \leq 2.4$, $F_{max} \leq 1.8$ and $S_{min} \leq 3.5$. The base-band noise power was estimated using the assumption: $A_{eq} = 3.49$.

The resulting parameters are $(t_1, t_2, t_3) = (0.922, -0.0588, 0.6347)$. In this case, all of the three constraints were active (i.e., the upper bounds were reached). The found parameters satisfy consequently three equations with three variables and the parameters are hence given by the constraints only.

The predicted σ_b^2 was -128.7 dB and the simulated was -130.5 dB. Fig. 7.13 shows a maximum transient length plot for this modulator (same simulation parameters as for Example 7.2). The plot indicates that the optimized modulator has a reliable MSA of approx. 0.42 and that the onset of instability occurs abruptly. A simulation with a ramp increasing with a rate as low as 10^{-8}/sample confirms this MSA. The modulator achieves consequently a dynamic range of approx. 120 dB, or almost 20 bits of resolution. This performance is comparable to a recently reported fifth order modulator [17]. The simulated output spectrum of the optimized modulator for a constant input of 1/64 is seen on Fig. 7.14 □

In the previous example, all three constraints were active. When similar optimizations are carried out on third order modulators with the same constraints, it is typically the F_{max} limitation that is the dominating constraint, while the S_{min} constraint is inactive.

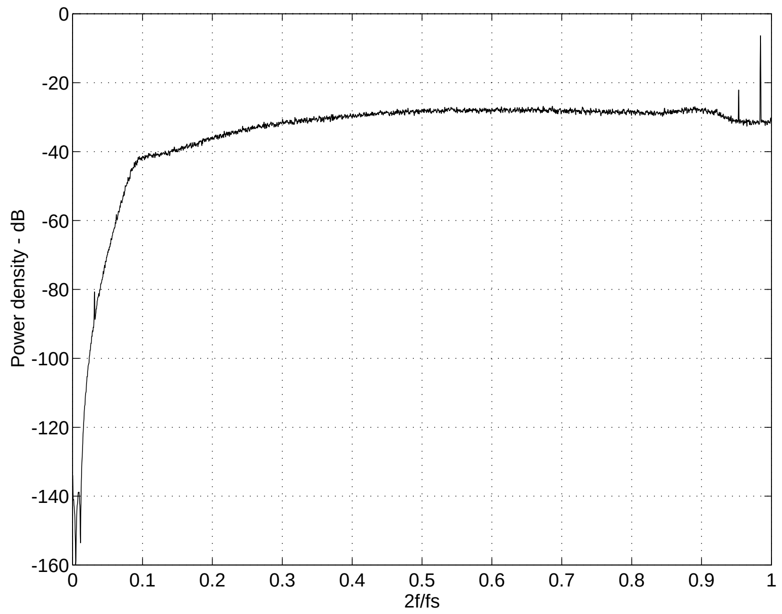

Figure 7.14: Estimated output power spectrum for the optimized modulator with the constraints $A_{min} \leq 2.4$, $F_{max} \leq 1.8$ and $S_{min} \leq 3.5$.

Conversely, for higher order optimizations ($N > 4$), the S_{min} constraint is dominating while the F_{max} constraint is typically inactive.

Example 7.9 An optimization was carried out for a third order modulator with the same f_b and constraints as in Example 7.8. The optimal feedback filter pole frequencies for third order are given by [58]:

$$f_i \in \left\{0,\ \pm\sqrt{\frac{3}{5}} \cdot f_b\right\} \tag{7.16}$$

The optimization gave the parameters $t_1 = 0.7752$ and $t_2 = -0.0663$. Only the $F_{max} \leq 1.8$ constraint was active: the optimization yielded $S_{min} = 3.22$ and $A_{min} = 2.32$. The predicted σ_b^2 was -96.7 dB and the simulated was -97.7 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.41.

An optimized fifth order modulator was designed with the same constraints. The optimal pole frequencies are [58]:

$$f_i = 0,\ \pm\sqrt{\tfrac{5}{9} \pm \sqrt{\left(\tfrac{5}{9}\right)^2 - \tfrac{5}{21}}} \cdot f_b \tag{7.17}$$

The optimization gave the parameters:

$$(t_1, \cdots t_4) = 0.9697, -0.0328, 0.8001, -0.0452)$$

Both the $S_{min} \leq 3.5$ and $A_{min} \leq 2.4$ constraints were active while F_{max} was 1.787. The predicted σ_b^2 was -144 dB and the simulated was -144.7 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.37. This corresponds an SNR of 133 dB for maximum amplitude sinusoidal input. A modulator for a commercial 64 times oversampling fifth-order A/D-converter was reported to yield a maximum simulated SNR of 119 dB for sinusoidal input with an amplitude of 0.4 [17]. The modulator has pole frequencies at 11kHz and 19kHz. However, the base-band used for the reported SNR figure was DC-22kHz with a sampling rate of $64 \cdot 48$ kHz and the SNR figure was actually measured with sinusoidal input. Even when these differences are compensated, the optimized modulator compares favorably to the reported commercial design.

For the sake of comparison, a fifth order modulator was designed using the NTF-prototype method described in Sec. 6.7. The feedback filter zeros were found using a Chebychev II prototype with Rs= 111.75 dB. The poles were given by Eq. (7.17). This modulator has $A_{min} = 2.4$, $S_{min} = 3.5$ and $F_{max} = 1.7$. The predicted σ_b^2 was -139.8 dB and the simulated was -140.97 dB. A simulation with a ramp increasing with a rate of 10^{-8}/sample yielded an MSA of 0.38. This modulator fulfills the constraints used for the optimization but the noise suppression is approx. 4 dB lower than for the optimized modulator
$\square$

7.7 Considerations for Higher Order Designs

All of the four designs presented in Example 7.8 and Example 7.9 fulfill the same set of constraints and these four designs work reliably within nearly the same amplitude range in agreement with the Gaussian criterion. This shows that the used constraints are consistent for orders at least up to five.

It is generally true that a better performance is obtained by raising the order of the modulator for a given oversampling ratio (or base-band). The noise shaping theorem of Th. 6.1 gives a very simple explanation: For stability reasons, the high-frequency NTF gain must be limited (e.g., by bounding A_{min} or S_{min}) and the only way to improve the base-band rejection without violating the noise shaping theorem is to reduce the transition band from base-band to pass-band. An efficient way to reduce the transition band is to increase the filter order. As the transition band is reduced, both the decimation filter in A/D converters and the analog post-filters in D/A converters must be more steep and, hence, more complex.

Even with a 'brick-wall' NTF-characteristic, the noise shaping theorem and the need for suitable bounds on the NTF high-frequency gain limit the attainable performance. This is quite comforting, since it would be strange if it was possible to design a modulator yielding more bits or resolution than the oversampling factor (see Ch. 1). The rest of this section demonstrates the performance limit for modulators intended for 32 times oversampling.

Example 7.9 showed that the NTF-prototype method using Chebychev II filters gives good and reliable fifth order modulators. However, the optimization procedure obtained a 4 dB better result with almost the same stability. Unfortunately, it turns out that the Chebychev II prototype method gives more and more unreliable modulators for increasing order.

Example 7.10 A modulator order of eight is necessary for a CD-quality converter operating at only 32 times oversampling. An eight order Chebychev II prototype with fb= $1/32 \cdot 20/24$ and Rs= 101.5 dB achieves a A_{min} value of 2.4. Unfortunately, the S_{min} value is as high as 4.73 and this indicates that the modulator is rather unreliable. The F_{max} value is as low as 1.66. A ramp with a sweep rate of 10^{-6}/sample gave a MSA of 0.4 while a rate of 10^{-9}/sample yielded 0.21. The measured base-band noise power σ_b^2 for a constant input of 1/64 was measured to -128.9 dB. □

The previous example showed that loop filter optimization becomes mandatory for very high order modulators. The NTF-prototype method represents a one-dimensional curve in a very high dimensional parameter space. It is obvious that improvements can be obtained by exploring the entire parameter space.

Until now, the signal transfer function (STF) has received no attention. The STF as given by Eq. (6.2) will for the usual sigma-delta modulator normally be flat in the base-band due to the high base-band gain of the feedback filter. This is not the case for the popular multiple feedback topology (see Example 2.3). This topology is equivalent to a generic modulator (see Fig. 2.4) with $H(z) = C(z)/D(z)$ and $G(z) = 1/D(z)$ [49]. This means that a multiple feedback modulator needs a prefilter with a transfer function approximately equal to $C(z)$ in order to equalize the base-band STF. For low order designs with a high oversampling ratio, such equalization is not necessary because $C(z)$ is typically flat in the base-band. For higher order systems with lower oversampling ratios, the base-bandwidth is comparable to the transition bandwidth of the noise transfer function and there will generally be untolerable base-band ripple.

The necessary equalization can be accomplished digitally for both A/D and D/A converters. In an A/D-converter, the equalization is performed digitally on the one-bit stream. Consequently, if the STF is peaking significantly in some frequency bands, the MSA as seen from the modulator input will be reduced accordingly. In the cases where equalization is not possible, the feedback filter optimization must include a constraint on the maximum allowable STF base-band ripple. Such a constraint can decrease the attainable signal-to-noise ratio considerably. The use of an STF-constraint was also used in [29].

When $C(z)$ is significantly non-flat in the base-band, the base-band behavior of the NTF will also be affected. Hence, the optimization of the zeros given by $C(z)$ and the poles given by $D(z)$ can theoretically no longer be carried out separately due to this interaction. On the other hand, the increased dimensionality of the unified optimization problem can be too impractical.

Example 7.11 The 8th order modulator for 32 times oversampling of Example 7.10 was optimized with the usual constraints: $A_{min} \leq 2.4$, $S_{min} \leq 3.5$ and $F_{max} \leq 1.8$. The feedback filter poles were chosen according to [58]. The STF in a multiple feedback configuration is peaking 10.4 dB at the base-band edge. The base-band noise power was measured to -129.3 dB, i.e., better than for the NTF-prototype design in Example 7.10. The optimized modulator is very reliable and the MSA was estimated to 0.41 using a ramp with a rate of 10^{-9}/sample.

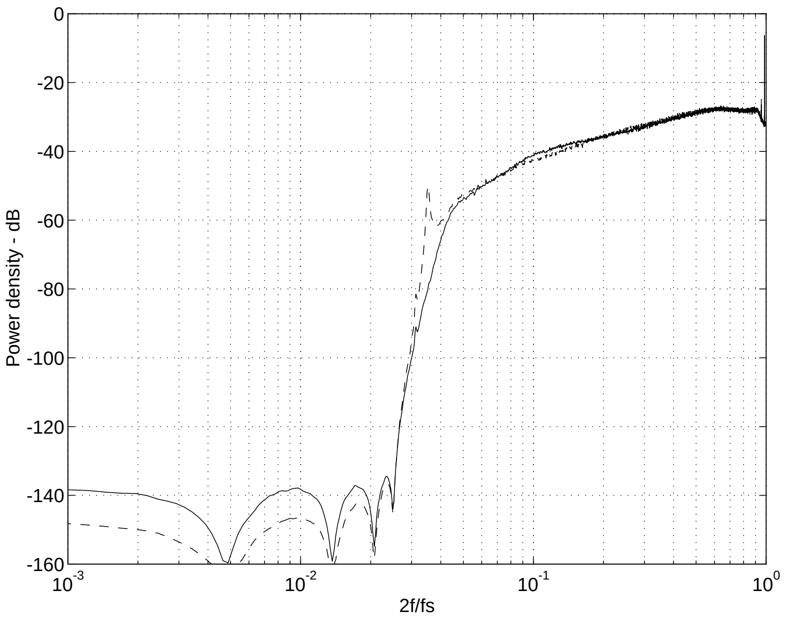

Figure 7.15: Estimated power spectra for the two optimized 8th order modulators in Example 7.11. The dotted line represents the modulator optimized without an STF magnitude constraint.

The modulator was reoptimized using a ± 0.25 dB constraint on the multiple feedback STF magnitude within the base-band. The resulting modulator yielded a σ_b^2 of -125.7 dB, i.e., the new STF constraint reduces the noise rejection. Using a ramp with a rate of 10^{-9}/sample, the MSA was estimated to 0.36.

Fig. 7.15 shows the estimated power spectra of the two modulators with a constant input of 1/64. Note that the STF peaking can be observed on the base-band spectrum for the modulator optimized without an STF constraint. This modulator shows also a noise spectrum increasing faster above the base-band than for the modulator with the STF constraint. In a D/A converter, it is extremely difficult to suppress such out-of-band noise using analog filters of reasonable complexity. □

The previous example showed that the A_{min} and S_{min} constraints are fairly consistent even for eighth order modulators. However, the quadratic programming used by the CONSTR routine is very sensitive for high-dimensional systems, i.e., the optimization routine might be unstable for some initial guesses. It is strongly recommended to impose tight bounds on the parameters in order to confine the optimization process.

7.8 Optimization of Unusual Topologies

An extension of the multiple feedback modulator topology has been suggested in [43]. This topology includes an extra FIR filter in the feedback section, cf. Fig. 7.16. The hardware penalty for such an FIR filter is quite low, since the filter operates with a one-bit signal

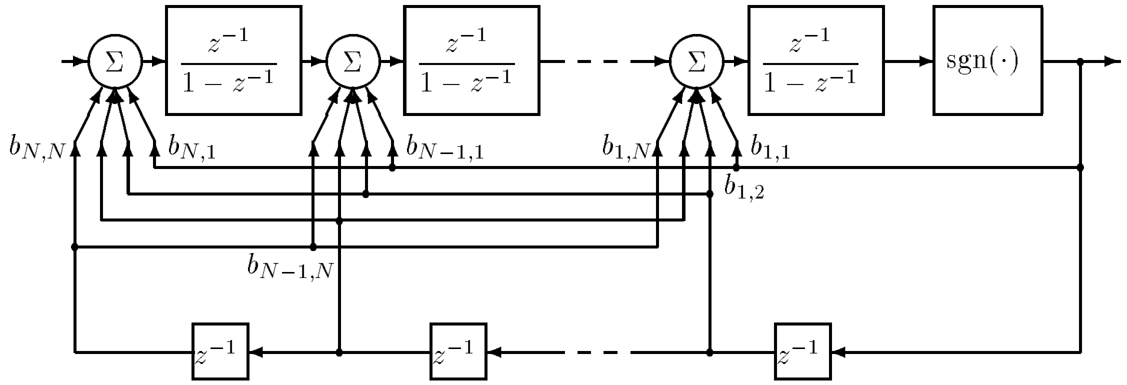

Figure 7.16: A proposed extension of the usual multiple feedback modulator (the figure is a modified version of a topology suggested in [43]). The topology allows an equivalent feedback filter with more zeros than poles.

as input. In a digital implementation, such FIR filter can be implemented using table look-up. In the analog domain, the structure on Fig. 7.16 can directly be used. The $b_{i,j}$-coefficients form a (N, N)-matrix which determine the $2N-1$ zeros of the feedback filter. This gives a redundancy which allows the use of simpler $b_{i,j}$ coefficients. A fourth order modulator designed with power-of-two coefficients was presented in [43].

The new topology allows an equivalent feedback filter with a numerator with a higher degree than the denominator, i.e., the number of design parameters increases. Such feedback filters can be split into a cascade of an FIR filter followed by a conventional IIR Nth order filter where N is the number of feedback filter poles. Feedback filters of this type can not be designed by the NTF-prototype method using conventional high-pass filters. Consequently, a direct design method like the proposed optimization framework must be used.

A number of experiments were carried out on a fourth order modulator, i.e., a modulator with four poles in the feedback filter. The modulator was optimized using 3, 4 and 5 feedback filter zeros. A number of 5 feedback zeros corresponds to an extra second order FIR filter. Only marginal improvements were obtained using these extra zeros: the base-band noise suppression was improved approx. 1 dB while the stability was preserved.

7.9 Summary

A framework for feedback filter optimization has been presented in this chapter. The method is based on a constrained optimization of a prediction of the base-band noise suppression. The base-band noise power prediction was based on the quasilinear modeling derived in Chapter 6. The constraints ensure that the modulator has a prescribed stable and reliable amplitude range. The modulator stability was analyzed in the BIBO sense using an equivalent noise-shaper topology. The scaling invariance of the feedback filter was utilized to improve the BIBO criterion. It was demonstrated that the BIBO-criterion can be based the minimum one-norm of the noise transfer function while the Gaussian criterion from Sec. 6.6is based on the minimum two-norm. A third stability measure was defined using the infinity norm of the signal transfer function.

Optimizations constrained by proper bounds on the three stability measures yielded good results. The constraints were consistent for a wide range of modulator orders, i.e., the same set of constraints yielded fairly consistent stability properties irrespectively of the modulator order.

Other constraints can easily be included. For instance, in some cases it is convenient to restrict the ripple of the base-band signal transfer function. Another possibility is to include the characteristic of a succeeding decimation filter into the optimization.

Chapter 8

Designing Tone Free Modulators

8.1 The Tone Problem

It is a well known problem that Σ-Δ modulators produce in-band as well as out-of-band tones in the presence of low amplitude constant input. In particular, there is often a very predominant tone near half the sample rate at the frequency $f = (1 - \mathrm{dc})f_s/2$ when the modulator is fed with constant input dc [2, 24, 32, 33]. This tone can be observed on all the spectra shown previously for modulators with constant input. The tone is actually frequency modulated by the input signal: sinusoidal input generates characteristic side-band tones (see the paper in Appendix C).

The persistence of this often very strong tone near half the sample rate is not easily explained. The classical first order modulator has the $\overline{10}$ limit cycle as the only steady-state solution for zero input [15]. This limit cycle gives indeed a strong tone at $f_s/2$. As a small dc-bias m_x is added to the input, the modulator will from time to time generate a code segment consisting of two identical codes. This happens because the error between modulator input and output accumulates in the integrator inside the loop. Consequently, the mean value of the modulator output must be equal to the dc-bias, i.e., $m_y = m_x$. The frequency of code repetitions must then be $m_x \cdot f_s$ and this also explains the tone at $(1 - m_x)f_s/2$. Higher order modulators seem to inherit this tone mechanism, though the output code becomes more complex. An Nth-order multiple feedback modulator can be perceived as a first order modulator embedded in a feedback loop containing $N - 1$ integrators. The extra integrators do only modify the low-frequency spectrum, i.e., more noise is removed from the base-band. The effect at higher frequencies is very little due to the low-pass filtering action of the extra integrators, and hence, the high-frequency tone caused by the innermost integrator is persisting [2].

The high-frequency tone is often accompanied by in-band tones at multiples of $f = \mathrm{dc}f_s$. This can be interpreted as a kind of intermodulation between the high-frequency tone and its image above half the sample rate. The high frequency tone is often very strong and can be quite close to full scale. This causes the random-like part of the quantization noise to be modulated, i.e., the total instantaneous output power is always unity and this must be shared by the random component and the tone. The result is that also the base-band noise power is modulated by the frequency $f = \mathrm{dc}f_s$, i.e., the random part of the quantization noise becomes non-stationary. The noise modulation itself cannot be seen on usual power spectrum estimates because time averaging is required. However it has been pointed out that periodic noise modulation can be perceived as a tone by the human ear [40]

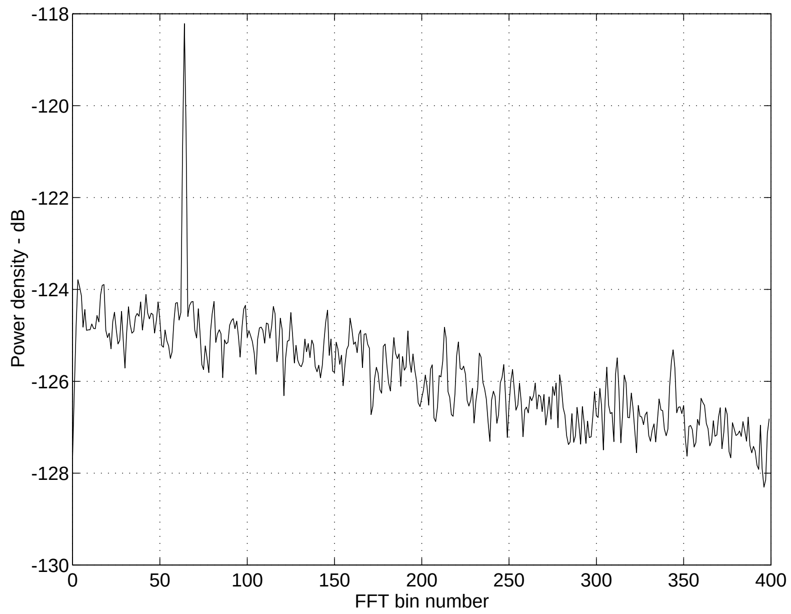

Figure 8.1: Estimated power spectrum of the decimated, high-pass filtered and squared modulator output based on 100.000 decimated samples simulated with a constant input of $1/(64 \cdot 128)$.

Example 8.1 The fourth order modulator of Example 7.8 was simulated with a constant input of $1/(64 \cdot 128)$ and the modulator output was decimated 64 times using a very selective filter. The resulting base-band noise signal contained a rather strong tone at the frequency $f = f_s/(64 \cdot 128)$ accompanied by harmonics. The decimated base-band signal was high-pass filtered in order to remove the dc-component, the tone and its harmonics. The filter used was an 8th order Chebychev II filter with Rs= 60 dB and stop-band edge at $0.1 \cdot f_s/2$. The filtered signal was subsequently squared and the power spectrum was estimated using the usual Welch method based on windowed 8k FFT's. The resulting spectrum is plotted in Fig. 8.1, and this spectrum shows a significant peak at bin number 64 which corresponds to the frequency $f = f_s/(64 \cdot 128)$. This peak is not caused by the tone itself due to the high-pass filtering. A possible interpretation is that the variance of the base-band signal is modulated periodically, i.e., the quantization noise is clearly non-stationary □

The simulations in Example 8.1 seem to confirm the periodic noise modulation is caused by the very strong high-frequency tone. There seems also to be a strong link between this tone and the base-band tones. Unfortunately, the out-of-band tones have not attracted very much attention in the literature so far.

Another observation from the time domain is that the out-of-band modulator output noise can be very 'impulsive' of nature, i.e., the signal power is concentrated in short bursts with high peak values [42].

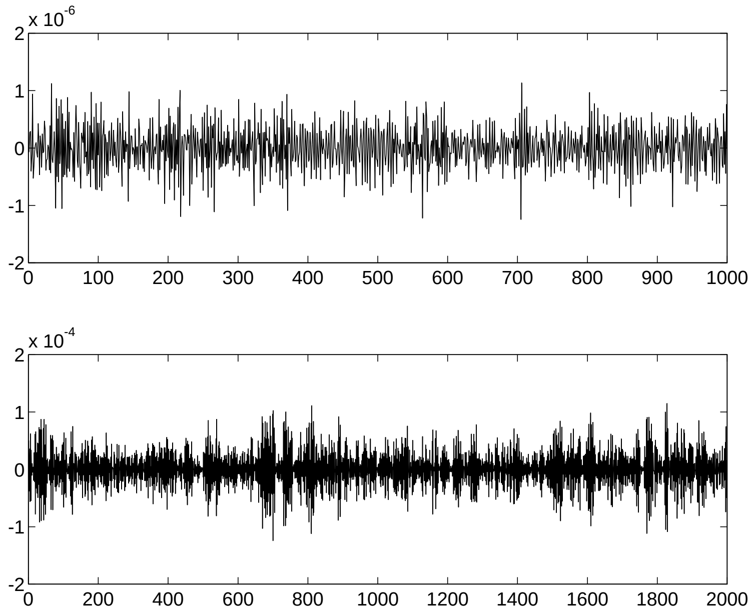

Figure 8.2: Decimated and high-pass filtered output signals of the fourth order modulator of Example 7.8 fed with a constant input of $1/(64 \cdot 512)$. The upper graph is decimated 64 times and the lower is decimated 32 times.

Example 8.2 Fig. 8.2 shows two signal sequences for the fourth order modulator from Example 7.8 fed with a constant input of $1/(64 \cdot 512)$. The upper graph shows the 64 times decimated and subsequently high-pass filtered modulator output. The plotted sequence corresponds to approx. two periods of the noise modulation which is only barely visible. The lower graph shows a similar sequence obtained by only 32 times decimation and subsequent high-pass filtering. Note the difference in magnitude scale due to the out-of-band noise. The length of the sequence matches the upper graph. The energy of the 32 times decimated sequence is, contrary to the 64 times decimated sequence, clearly concentrated in short bursts. This kind of non-stationary behavior is not revealed by usual spectral analysis □

The presence of tones and noise modulation are both unwanted effects of a modulation scheme. Even if the in-band tones are suppressed, the high frequency tone will still cause noise modulation. Furthermore, the presence of analog circuit non-linearities can still produce in-band tones and harmonic distortion due to intermodulation effects [2]. Consequently, the out-of-band tones must be taken into consideration when data converters are designed and evaluated.

8.2 Methods for Tone Suppression

The tone problem as far as in-band tones are concerned has received a lot of attention and several methods have been proposed for suppression of the tones [26, 37, 40, 57]. It

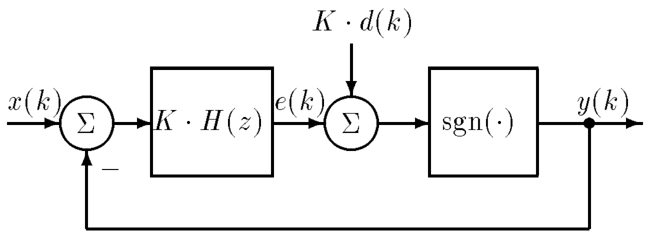

Figure 8.3: Σ-Δ modulator with scaled feedback filter $H(z)$ and dither source $d(k)$.

is generally believed that higher order modulators are less tonal [17]. This is only true when the in-band tone components are taken into consideration: the ratio between the power of then in-band tones and the in-band noise power gets typically lower for higher order modulation. However, the strength of the high-frequency tone changes only very little, i.e., the amplitude is generally in the range from -10 dBFs to -5 dBFs [1]. In many cases, the tone is stronger than the maximum amplitude of the sinusoidal modulator input. Consequently, high order modulators will still suffer from noise modulation and in-band tones caused by intermodulation due to circuit non-linearities.

A common approach is to add dither noise to the modulator loop [12, 40, 41, 42], e.g., accomplished by adding noise to the modulator input. However, this requires that the noise is heavily high-pass filtered in order not to sacrifice the signal-to-noise ratio. A more convenient approach is to add the dither to the quantizer input as shown in Fig. 8.3. This will automatically shape the spectral contribution of the dither according to the noise transfer function. The two approaches are theoretically equivalent when a suitable dither pre-filtering is applied in one of the cases.

The addition of a dither noise source has the effect that possible tones are 'dissolved', i.e., the signal power of the tones is partly transformed into random noise. At some dither level, the tones will be 'burried' in the random noise floor of a power spectrum estimate.

The effect of a dither source is 'discrete', i.e., the output code $y(k)$ will now and then change sign due to the dither signal. Theoretically, there is a large class of neutral dither signals which will not change the output code, e.g., as long as $d(k)$ has the same sign as $e(k)$, the operation of the modulator will be unchanged even for very large amplitude dither.

The presence of a dither source prior to the quantizer input will generally reduce the correlation between $e(k)$ and $y(k)$ and, hence, the quantizer gain K given by the quasilinear model will be reduced. This reduces the noise amplification factor A according to the A(K) curve of the feedback filter, i.e., the quantization noise power σ_q^2 of the quantizer is increased. Consequently, the reduced loop gain and the increased σ_q^2 will increase the base-band noise power σ_b^2. As the quantizer gain K is reduced, the equilibrium will move towards the A_{min}-point, i.e., the stability is deteriorated; hence, the stable amplitude range is reduced. The same conclusion can be drawn from the BIBO considerations in Sec. 7.3:

[1] dB relative to a full-scale sinusoid

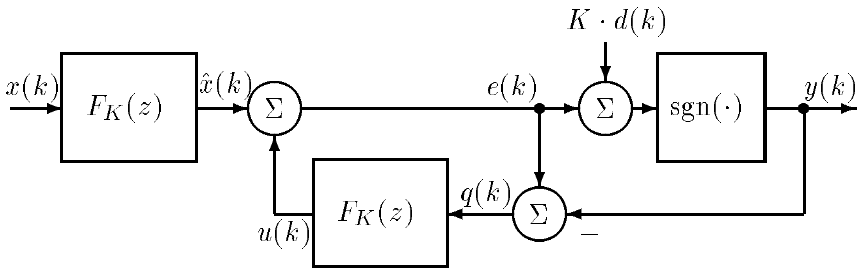

Figure 8.4: Equivalent noise-shaper topology of a Σ-Δ modulator with scaled feedback filter and dither source (see Fig. 8.3. The filter $F_K(z)$ is given by Eq. (7.7).

the magnitude of the quantization noise $q(k)$ can be as large as $1 + \|d(k)\|_\infty$ where $d(k)$ is the dither sequence (see the equivalent noise-shaper in Fig. 8.4). This maximum value occurs, e.g., when $e(k) = \|d(k)\|_\infty$ and $d(k) = -\|d(k)\|_\infty$. A BIBO stability criterion can now be derived:

$$\|f(k)\|_1(1 + \|d(k)\|_\infty) \leq 2 - \|\hat{x}(k)\|_\infty \tag{8.1}$$

where $\|f(k)\|$ is the one-norm of the impulse response $f(k)$ associated with $F(z)$.

A similar result was derived in [41].

The feedback filter scaling invariance applies only when the dither signal is scaled accordingly. Hence, a specification of the dither span without a specification of the scaling of the feedback filter is imprecise. Utilizing a feedback filter and dither scaling by a factor K, the BIBO criterion becomes:

$$(S(K) - 1)(1 + K\|d(k)\|_\infty) \leq 2 - \|\hat{x}(k)\|_\infty \tag{8.2}$$

where $S(K) = \|f_K(k)\|_1 + 1$ is the one-norm of the impulse response of the noise transfer function as a function of the quantizer gain K, cf. Eq. (7.9).

The best (i.e., the most optimistic) criterion is obtained for the K-value that minimizes the left hand side of Eq. (8.2). Eq. (8.2) shows that as the dither magnitude $\|d(k)\|_\infty$ is increased, the stable amplitude range is reduced.

Consequently, both the quailinear analysis and the BIBO criterion show that modulator feedback filters designed for use with a dither noise source must be designed very conservatively in order to preserve a prescribed stable amplitude range.

As an alternative to dither, it has been suggested that the use of chaotic modulators can suppress the tones [26, 37, 41, 57]. The output of a chaotic system is generally non-periodic and the spectrum must be continuous. However, this property does not exclude that tones might exist combined with non-periodic noise like components. As stated in Chapter 3, chaotic modulators are obtained when one or more of the poles of the feedback filter are located outside the unit circle. This means that the noise transfer function is mixed phase, i.e., a minimum phase transfer function with the same amplitude characteristic can be obtained by reflecting the zeros inside the unit circle to their reciprocal locations. The zero reflected minimum phase noise transfer function has a lower two-norm or noise amplification factor A than the mixed phase characteristic due to the ntf(0) = 1 scaling requirement. The reason is that any mixed phase filter can be composed of a minimum phase part and an all-pass part [44]. Due to the necessary scaling, the amplitude gain of the all-pass term is greater than unity. The zero reflected minimum phase filter is obtained when the all-pass term is removed. Consequently, for a given base-band noise suppression, a chaotic modulator will always have a larger A_{min}-value than for a non-chaotic modulator. According to the Gaussian criterion, this means that the stable amplitude range is reduced. A similar argument holds for the BIBO-criterion since the one-norm of the mixed phase noise transfer function is higher than for the minimum phase. Consequently, both the use of dither and chaos will deteriorate the loop stability.

The paper in Appendix A compares dithering and the use of chaos with respect to the suppression of the high-frequency tones. The comparisons are made for sixth order modulators intended for 64 times oversampled CD-standard digital audio reproduction. The conclusion is that both dithering and chaos can attenuate the high-frequency tones more than 30 dB. The price paid is a significant (20-30 dB) reduction of the attainable signal-to-noise ratio. It was shown that a feasible way to design tone free chaotic modulators is to extend a conventional NTF-prototype with a first-order all-pass term having a (real valued) zero α outside the unit circle. The resulting feedback filter will then have a pole at α. The best result is obtained when α is negative, i.e., closer to the tone frequencies near $z = -1$ which corresponds to half the sample rate. It was concluded that a tonefree and dithered sixth order modulator has a performance similar to a tonefree and chaotic seventh order modulator with a sixth order NTF magnitude characteristic. This shows that the use of chaos and dithering have very similar effects on the tones and the stability.

A totally different approach for tone suppression is used in [33]: a fine-scale 'trim bit' with, e.g., 8-bit lower significance is added to the coarse one-bit quantizer output and the result is a rather unusual 4-level quantizer with a large number of thresholds where the output code changes. It is demonstrated that such coarse/fine quantizer gives better low level resolution without in-band tones. However, the effect on the strong out-of-band tones is not clear.

8.3 Measures of Information Loss

Both the use of dithering and chaos introduce an element of uncertainty into the modulation. For a dithered modulator, the next output code will to some extend be uncertain. If the state of a chaotic system is known with some precision at time-step k, then the state of the system at time-step $k + 1$ is known with a reduced precision due to the sensitivity to initial conditions. This property represents a loss of information for each time-step.

The product of the feedback filter pole moduli was introduced as the system map expansion factor Λ, cf. Eq. (4.7). The rate of information loss can be expressed in bits per time-step as the base two logarithm of Λ. As Λ approaches two, the loss of information becomes close to one bit per time-step. It was previously shown that a first order modulator becomes unstable when Λ is greater than two (see Example 3.1). Higher order systems become typically unstable for lower Λ values. Consequently, an information loss rate of one-bit per time-step is the upper limit for a one-bit modulator. The noise shaping theorem (Theorem 6.1) is in fact inspired by information capacity considerations. When the equality of Eq. (6.4) holds, the information capacity of 'the noise shaped transmission channel' is preserved [20]. This is only the case for minimum phase noise transfer functions, i.e., for non-chaotic modulators. Consequently, the the use of chaos reduces the information capacity.

Based on listening experience, it has been claimed in [31] that multiple feedback filter poles at $z = 1$ gives less correlated quantization noise than for complex unit circle poles. A likely explanation is that the number of periodic points increases as $n^{(N-1)}$ where n is the period length and N is the multiplicity of the $z = 1$ poles (i.e., the order), whereas the number of periodic points is asymptotically constant for the case of complex unit circle poles.

It is also possible to quantify the loss of information for a dithered modulator. Let $X = \mathrm{sgn}(e(k))$ and $Y = \mathrm{sgn}(e(k) + d(k))$ be stochastic variables. The conditional entropy $\mathrm{H}(Y|X)$ (in bits) expresses the uncertainty about the current modulator output code Y given the output code X without dither. The undithered system yields $\mathrm{H}(Y|X) = 0$, since there is no uncertainty about Y once X is known. As more and more dither is added, $\mathrm{H}(Y|X)$ increases. The upper limit is of course one-bit in the extreme case where Y is statistically independent of X and the two outcomes of Y are equally likely.

It is now possible to compare the amounts of dither and chaos. The dithered sixth order example from the paper in Appendix A operates for zero input with an uncertainty $\mathrm{H}(Y|X)$ of approx. 0.69 bit/time-step. The all-pass derived sixth order modulator has a 'chaotic' pole at $\alpha = -1.25$ corresponding to an information loss rate of 0.32 bits/time-step. Consequently, the chaotic modulator can suppress the high-frequency tone at a lower rate of information loss than for the dithered example. If the all-pass zero is flipped to $\alpha = 1.25$, the high-frequency tone is only reduced slightly while the pole modulo is preserved.

8.4 Limit Cycle Analysis of Dithered Modulators

The limit cycle analysis framework presented for chaotic modulators (Ch. 4) can be extended to include dithered modulators. The analysis uses the usual modulator topology of Fig. 8.3. To test if a given periodic code sequence exists as a limit cycle, the loop is cut open at the input of the quantizer and the periodic steady-state loop filter output $e(k)$ is found when the period code sequence $y(k)$ is fed back. If the quantizer will reproduce the output code sequence from $e(k)$, then the code sequence exists as a limit cycle. However, it should be emphasized that a modulator dithered by an independent, stochastic dither signal is not a deterministic autonomous system. Consequently, the concept of limit cycles does not exist for such dithered modulators. Only when the dither signal is a known periodic sequence, the limit cycle concept is meaningful.

The addition of a periodic dither signal $d(k)$ to $e(k)$ changes the composition of possible limit cycles, i.e., the dither can change the quantizer output code at suitable instants and other periodic sequences can be reproduced. This kind of limit cycle analysis provides information about the possible code sequences of a dithered modulator. A stochastic dither generator may theoretically produce a large class of periodic signals. Hence, the objective of the limit cycle analysis is to find every theoretically possible limit cycle for a given dither generator. In practise, the limit cycles are not observable for real simulations with random dither, since each of the limit cycles requires a particular class of periodic dither signal added in order to be sustained. This is exactly the same situation as for chaotic modulators: there is a large number of unstable limit cycles and, hence, it is very unlikely that a particular cycle is sustained for a longer period of time due the sensitivity to initial conditions. However, the knowledge of the composition of possible limit cycles describes the possible code sequences, i.e., the modulator will always be arbitrarily close to a limit cycle.

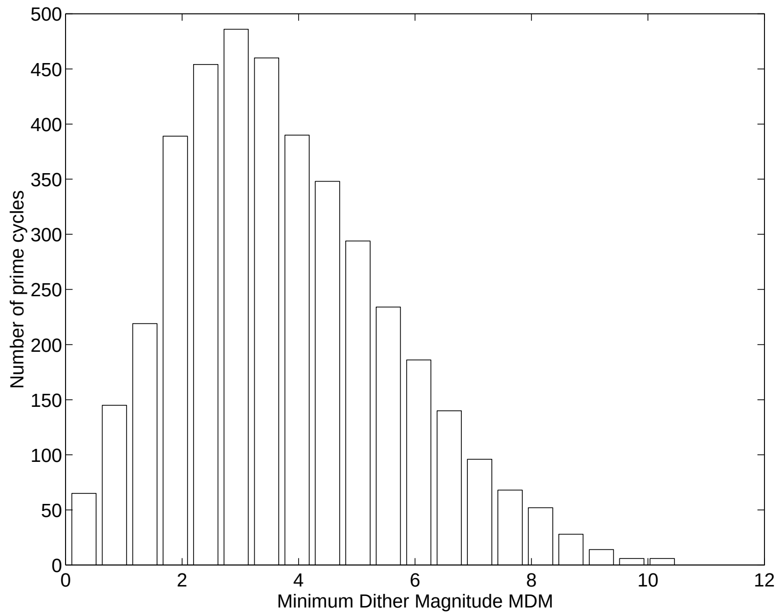

Figure 8.5: An Minimum dither magnitude (MDM) histogram for the set of period 16 prime cycles of the modulator with feedback filter cf. Eq. (8.4) with zero input. A total of 4080 prime cycles of length 16 exist cf. Table 4.1.

It was stated in Sec. 4.3 that the number of periodic points of a chaotic modulator must grow exponentially with the period length in order to maintain stability. To be more specific, the number of periodic points must grow like Λ^n where n is the period length and Λ is the expansion factor (i.e., the product of pole moduli greater than unity).

The limit cycle analysis can thus describe the similarity between the effects of dithering and the use of chaos. In both cases, the number of limit cycles increases and this tends to randomize the output code sequence.

It must be emphasized that the limit cycle analysis of a dithered modulator does not provide information about the stability, i.e., a modulator will always be less stable when random and independent dither is applied.

A practical way to analyze limit cycles for a dithered modulator is to find the steady state $e(k)$ signals for each of the prime cycles up to a given length. Subsequently, the minimum dither magnitude (MDM) necessary to reproduce the corresponding prime cycle is computed for each $e(k)$ sequence:

$$\mathrm{MDM}_i = -\min(e_i(k) \cdot y_i(k)) \tag{8.3}$$

where $y_i(k)$ and $e_i(k)$ are the code sequence and steady-state loop filter output corresponding to prime cycle i.

A negative MDM_i means that the prime cycle i exists without any dither, i.e., $y(k) = \mathrm{sgn}(e(k))$. If the dither signal $d(k)$ is a bounded random i.i.d. sequence then every prime cycle i satisfying $\mathrm{MDM}_i < \|d(k)\|_\infty$ will theoretically exist, i.e., there is a (small) possibility that the random dither signal will support the limit cycle for some time given the right initial conditions.

Example 8.3 This example investigates the second-order modulator feedback filter:

$$H(z) = \frac{z^{-1} - 0.5z^{-2}}{1 - 1.95z^{-1} + z^{-2}} \tag{8.4}$$

The feedback filter has complex unit circle poles, i.e., the number of periodic points is expected to be asymptotically constant without dither applied. The following analysis is accomplished for zero input.

Fig. 8.5 shows a histogram of the MDM-values of the set of period 16 prime cycles. There are no negative MDM-values, i.e., no period 16 prime cycles exist without dither. As the dither span is increased, the number of existing prime cycles increases. For a maximum dither magnitude of 10.5, every prime cycle of length 16 exists (a total of 4080, cf. Table 4.1).

Fig. 8.6 shows the number of periodic points versus the period length for different maximum dither magnitudes. The plot shows only values for even period length since there are generally fewer limit cycles with uneven length (these cycles have a non-zero mean value). For no dither, the number of period points levels out to a fairly constant number. For higher dither levels, the curves follows the same initial steep slope and at some point, the slope levels out. Unfortunately, the developed routines in MATLAB were not able to handle prime cycles of length greater than 16. However, from the plots on Fig. 8.6 it seems likely that the curves will continue to grow exponentially with exponents depending on the dither level, i.e., the same asymptotic behavior as for chaotic modulators □

The limit cycle analysis for a dithered modulators confirms the similarity between dithering an the use of chaos: In both cases, the number of periodic points increases (exponentially) with the period length.

A dithered non-chaotic modulator with poles inside the unit circle will typically display a dither threshold effect, i.e., for low dither levels, the output will be strictly periodic and at a certain dither level the output becomes more random. This phenomenon is easily explained: modulators with poles inside the unit circle have attracting limit cycles and if the dither level is insufficient to eventually change an output code, the limit cycle will trap the modulator. The limit cycle which can tolerate the highest dither levels will then dominate and asymptotically attract the system. This limit cycle has the lowest (i.e. most negative) MDM-value. Typically, the classical $\overline{10}$ limit cycle requires the most dither to knock out. This gives also a credible explanation on the persistence of the high-frequency tone.

Example 8.4 A modulator with a feedback filter given by Eq. (8.4) was modified such that the unit circle poles were scaled by a factor of 0.95. the steady state feedback filter output for the $\overline{10}$ limit cycle is $e(k) = (0.3995, -0.3995...)$, i.e., it requires a dither signal with a peak magnitude of at least 0.3995 to dissolve this attracting limit cycle. Simulations with uniformly distributed dither confirms this threshold effect □

For a non-chaotic modulator, it is meaningless to apply less dither than prescribed by the dither threshold.

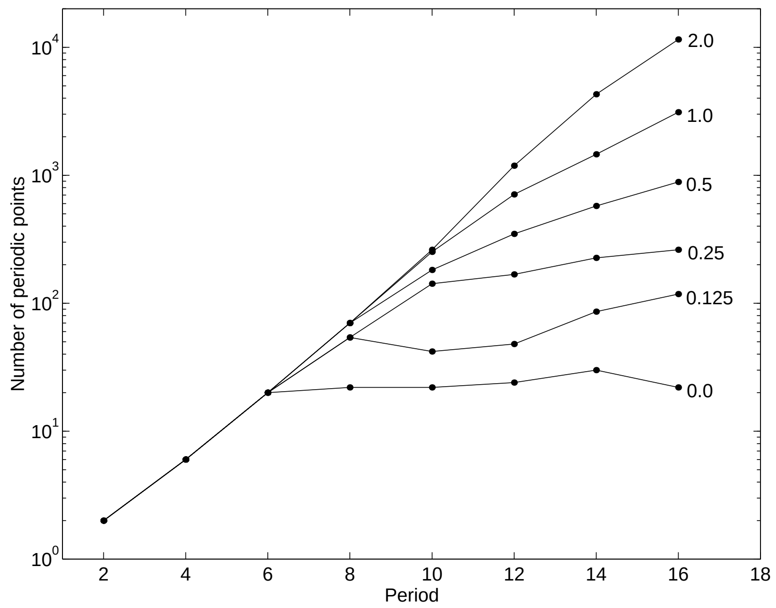

Figure 8.6: Number of periodic points versus period length for the modulator cf. Eq. (8.4) with zero input. The parameter on each plot is the peak dither magnitude $\|d(k)\|_\infty$.

8.5 Summary

This chapter reviewed different approaches for suppression of the strong tones which can be observed in the output spectra of modulators fed with small amplitude constant input. The paper in appendix A presented some practical design examples which demonstrated that both dithering and the use of chaos can suppress such tones. The price paid in both cases is a significantly reduced attainable signal-to-noise ratio due to a detrimental effect on the modulator stability.

The use of chaos and dithering was compared from a theoretical point of view. Both methods induce a degree of uncertainty into the modulation which can be measured as a rate of information loss. Another way of comparing was to investigate the number of periodic points versus the period length. Both dithering and chaos increases the number of possible limit cycles and this randomizes the output sequence.

The question of whether to use dithering or chaos is quite complex. In digital systems, pseudo random number generators can easily be made with a controllable probability distribution. On the other hand, it also takes only relatively little extra hardware to increase the order of the modulator by one as required by the first order all-pass extension.

In analog modulators, the design of dither noise generators with prescribed probability density functions can be difficult: noise generators based on thermal noise has Gaussian probability density and the variance is strongly temperature dependent. Furthermore, the use of thermal noise requires usually high amplification and this may lead to unintentional feedback and deteriorated power supply rejection. A better approach is to use a digital noise generator and a suitable D/A-converter. However, this approach is rather hardware consuming. Consequently, for A/D-converters, the use of chaos can be a very attractive.

Chapter 9

Σ-Δ Modulation using Vector Quantization

9.1 Introduction

In this chapter, a novel class of modulators is introduced. In the traditional Σ-Δ modulator, the output code is determined as the sign of the feedback filter output, i.e., the decision of which output code to generate is based on a one-dimensional projection (many-to-one mapping) of the general N-dimensional state-space of the feedback filter. From a theoretical point of view, such a one-dimensional projection does only convey very little information about the filter state. A more general way to make the output code decision is to use the whole state-space, i.e., the scalar one-bit quantizer is replaced by a one-bit *vector quantizer* (VQ) [19] which maps the entire state-space into the binary output alphabet. The class of modulators using vector quantization is naturally a superset of the traditional class of modulators with scalar quantization.

Each of the modulators using vector quantization can be transformed into an equivalent traditional modulator with a scalar quantizer operating on a scalar filter output $e(k)$ combined with a vector quantizer which can change the decision of the scalar quantizer. The negated transfer function from the modulator output to the scalar signal $e(k)$ is still defined as the feedback filter $H(z)$. Such transformation can naturally be accomplished in many ways. Fig. 9.1 shows the topology of the the proposed modulator with combined scalar and vector quantization.

Conceptually, the topology of Fig. 9.1 can also be perceived as a traditional modulator

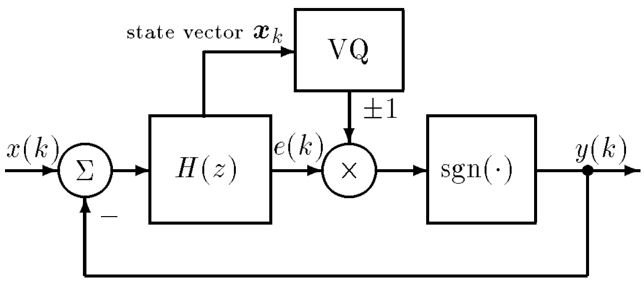

Figure 9.1: Σ-Δ modulator with vector quantization

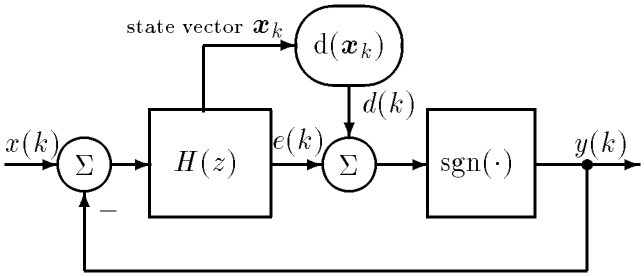

Figure 9.2: Σ-Δ modulator with vector quantization implemented as a deterministic dither signal

with a 'dither signal' added which is a deterministic projection of the filter state. Such topology is shown on Fig. 9.2. The dither signal produced by the generally nonlinear mapping $\mathrm{d}(\boldsymbol{x}_k)$ causes the output code of the scalar quantizer to change at some instants. If the dither signal is generated as a linear projection of the filter state, the system is equivalent to a traditional modulator with a modified $H(z)$ transfer function. To be more specific, a linear state-space projection can only change the zeros of $H(z)$ since the poles are always the eigenvalues of the transition matrix of the filter. Hence, the only interesting situation is when the deterministic dither signal is a nonlinear projection of the filter state. Since the usual scalar filter output $e(k)$ and the artificial dither signal is added before the sum is quantized by the scalar one-bit quantizer, the proposed new class of modulators could also be characterized as a modulator having a nonlinear feedback filter output. The nonlinear mapping could be implemented using a suitable neural network. It has previously been suggested to use a nonlinear feedback filter [65]; however, the proposed topology is not equivalent to a modulator with a vector quantizer

The question is obviously: is it possible to improve the performance of a modulator by the introduction of a vector quantizer? It seems very likely that it is possible to improve the stability. The reason is that a vector quantizer can allow more limit cycles to exist and this can be done in a selective way, i.e., the VQ can be designed to generate more limit cycles belonging to the bounded limit set. Recall from Ch. 2 and Ch. 4 that a chaotic limit set is made up of a large number of limit cycles and that the stability of the limit sets depends on the 'density' of limit cycles. The VQ might thus keep the modulator operating in a bounded region while the usual noise shaping effect is preserved, i.e., the base-band noise rejection is primarily controlled by the filter poles which remain the same. Such effects can increase the overall modulator performance.

The possibility of designing the composition of limit cycles might also be used to change and hopefully reduce the amplitude of the possible tones (see Ch. 8 for a discussion on tones). An unwanted limit cycle can be removed by a slight modification of the VQ. Since the modulator state is always arbitrarily close a limit cycle, the spectral properties of the modulator output is also influenced when changing the set of limit cycles. The possibility of using VQ for tone suppression is an open field for future research.

9.2 Specification of the VQ

It is impossible to investigate the entire class of modulators using vector quantization. Furthermore, there is currently no way to accurately model these systems and this makes systematic optimization very difficult. However, a systematic method for the design of the vector quantizer is proposed in [52] reprinted in Appendix B. The method starts from a usual modulator with loop filter $H(z)$ given in a state-space description. Subsequently, a so-called *legal set* is specified. The purpose of the vector quantizer cf. Fig. 9.1 is to ensure that the modulator never leaves the legal set. This is accomplished by identification of a number of regions F_n in the state space where the VQ should change the code generated by the scalar quantizer. These regions have the property that states in F_n leaves the legal set after precisely n time steps, if the usual quantizer is used. In addition, if a time-step is taken with an inverted output code followed by $n-1$ time steps with normal quantization, the modulator stays within the legal set. A direct implication of these properties is that the F_n regions are mutually disjoint. A VQ which changes the output when the modulator state is in one of the F_n regions for n below some limit, can help to confine the modulator operation (i.e., the limit set) to be inside the specified legal set. If the unmodified modulator was unstable, then the VQ can in some cases stabilize the system.

The proposed method requires some tweaking, i.e., the choice of the legal set is very crucial. A simple way to define the legal set is to specify an upper and a lower bound for the scalar filter output $e(k)$. This choice of legal set simplifies also the structure of the F_n sets, i.e., these sets are *polytopes* (sets delimited by hyper-planes in the state-space [19]). This enables the vector quantizer to be implemented using a number of scalar quantizers operating on auxiliary scalar filter outputs.

9.3 The Back-Step Algorithm

For a given legal set, the method presented in Appendix B defines a vector quantizer and a numerical test is presented which can tell if the modulator output should be inverted. However, this test has a high numerical complexity. In some cases, it is more efficient to use the so-called *back-step* algorithm: the modulator is simulated in forward time until the system leaves the legal set. At such instants, the algorithm steps back one step at the time and tries to flip the output code. Each time a code is flipped, the system is simulated in forward time until the time instant where the first legal set violation was encountered. The output codes are flipped successively back in time until the system stays within the legal set. A convenient way to arrange the numerical data is to store the signal sequences in circular buffers.

1. Simulate in forward time until the legal set is exceeded and let $\boldsymbol{n}_c$ be the time-step where this situation occurred.

2. **backstep** $= 0$

3. **backstep** $=$ **backstep** $+ 1$, if **backstep** $> n_{max}$ then goto 1.

4. Flip the output code at time time $\boldsymbol{n}_c-$**backstep** and simulate to time $\boldsymbol{n}_c$. If the system leaves the legal set then goto 3.

5. Goto 1

The back-step algorithm is faster when the vector quantizer only intervenes rarely, i.e., the algorithm stops only when leaving the legal set.

Example 9.1 An eighth-order modulator was designed using a Chebychev II prototype with the parameters Rs= 108 dB and fb= $1/32 \cdot 20/24$. The resulting feedback filter has a A_{min} of 2.63, and $S_{min} = 5.05$, i.e., the modulator is expected to be fairly unstable and unreliable (see Ch. 7).

The back-step algorithm was used on the eight-order feedback filter and a simulation was performed with a constant input of 1/256. An upper magnitude limit of $e(k)$ was set to 1.065. This resulted in a high frequency of output code changes. The output spectrum is seen on Fig. 9.3. The high vector quantization activity gives a significant suppression of the high frequency tone: the amplitude is as low as -17.7 dB. The base-band noise power is -131.1 dB. This shows that a reduction of the tones not necessarily sacrifices the in-band noise rejection.

The back-step algorithm was enable to stabilize the modulator for constant inputs up to 0.4, i.e., the vector quantization scheme can indeed extend the stable amplitude range considerably. In order to be stable, the $e(k)$ bounds were changed to -1.1 and 1.65. The algorithm had to go more steps back in time in order to limit the magnitude of $e(k)$. The spectrum showed again that the amplitude of high frequency tone was below -20 dB □

9.4 summary

A novel class of modulators employing one-bit vector quantization was introduced. This enables the output codes to be generated on the basis of the entire state-space information instead of on just a linear projection. Conceptually, such a modulator is equivalent to a modulator with a usual scalar quantizer which from time to time is overruled by a vector quantizer. Alternatively, this effect can also be achieved by a usual modulator with an added dither signal which is a deterministic (and nonlinear) function of the filter state. The operation of the feedback filter is unchanged, i.e., the filter has still the difference between the modulator input and output as input.

The introduction of a vector quantizer can change the stability and limit cycle composition of a modulator. A design method which aims at enhancing the stability was presented. A simulation example showed that the stability of an eighth-order modulator was enhanced significantly while the high frequency tone was reduced.

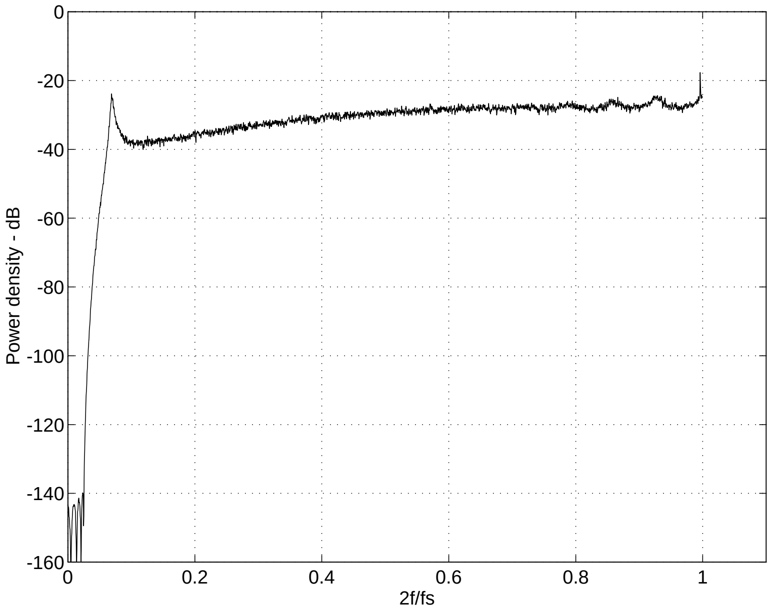

Figure 9.3: Simulated power spectrum estimate (500.000 samples) for an eighth-order modulator obtained from a Chebychev II prototype with the parameters Rs$= 108\,\mathrm{dB}$ and fb$= 1/32 \cdot 20/24$. The modulator was simulated using the back-step algorithm with a $|e(k)| \leq 1.065$ bound and a constant input of $1/256$.

Chapter 10

Conclusions

The present work has mainly focused on the stability of Σ-Δ modulators. The stability issue is the most delicate problem when design of high-order modulators is addressed. The analysis from the nonlinear dynamics point of view, as presented in part I, explained the onset of instability in a chaotic modulator as a boundary crisis emerging when the limit set (attractor) of the system collides with the boundary of its basin of attraction. The degree of instability can be quantified using the escape rate.

Even for second-order systems, the basins of attraction revealed extreme diversity and complexity. This fact tells that a simple and general stability criterion is very difficult to find: Stability can be proven by showing the existence of a bounded trapping region which naturally must be inside the basin of attraction. In the simple first-order case, such trapping region is an interval. However, as the dimension increases, more and more complex geometric analysis has to be applied. If a given stability criterion proves the existence of a simple class (e.g., hyper cubes) of trapping regions, then the criterion will be very conservative, especially for high-order systems. This is in fact the case for the BIBO criterion based on the NTF one-norm, as presented in Sec. 7.3. Less conservative stability criteria are of higher complexity (i.e., proving more complicated sets as trapping regions) — in the limit of complexity, it is just as easy to just simulate the system to find the answer.

The Gaussian stability criterion as presented in Ch. 6 takes advantage of the regularity hidden in the complexity of high-order systems. The Gaussian criterion is based on considerations for the power of the signals within the modulator loop, i.e., the criterion is based on two-norms. Unfortunately, this criterion is only approximate, but on the other hand, it is generally fairly precise for reasonably 'well behaved' systems.

Empirical stability studies showed that some modulators have very unreliable behavior giving a very slow transition to the unstable regime. Unreliable modulators may also have discrete constant input values which causes instability. The overwhelming problem concerned with unreliable modulators is the existence of regions or discrete values of constant input where the escape rate can be extremely small. This means that simulations cannot be trusted since these modulators are seemingly stable for most inputs.

Ch. 7 combined the approximate Gaussian stability criterion with the rigorous one-norm based BIBO criterion and an infinity-norm based stability measure. The result was a set of constraints which consistently lead to very reliable modulators in connection with an optimization routine. The resulting modulators had a very sharp transition to instability and this means that the maximum stable amplitude limit found by simulations can be trusted, i.e., the modulators are expected to be very stable if the input is magnitude limited slightly below the found amplitude limit. The results were used to implement a highly stable eighth-order Σ-Δ DAC.

The topic of stability was also investigated from a more abstract mathematical point of view, i.e., the symbolic dynamics analysis presented in Ch. 4. The conclusion is that for stable system, the growth in the number of periodic points must outbalance the volume expansion of the system map of the modulator. The results are not suited for practical design, and only in a few extreme cases, it was possible to derive an analytical expression for the escape rate. However, the context of symbolic dynamics gives an enlightening new view on the stability issue.

The symbolic dynamics was also used to show the similarities between dithering and the use of chaos. In both cases, the number of periodic points increases with the period length.

In Ch. 9 a new class of modulators using vector quantization was presented. It was shown that a suitably designed vector quantizer can improve stability considerably. The symbolic dynamics can explain this: The vector quantizer might change the composition of limit cycles, and this might increase the growth of periodic points.

The existence of spurious tones in the presence of small dc input bias is another problem concerned with Σ-Δ modulators. The tone problem was treated empirically, i.e., no theoretical analysis was presented. The use of chaos and dithering was compared and the result was that both options are nearly equally efficient, but there is a severe SNR penalty. In Ch. 9 it was shown that heavy use of vector quantization might both improve the stability and reduce the amplitude of the tones without any SNR penalty. This fact should obviously be subject to future research.

It is still an open question why spurious tones might persist even high amounts of dither or chaos. Again, more theoretical analysis is needed.

Appendix A

Appendix for Chapter 8

This appendix is connected to chapter 8 and is a preliminary preprint of the paper: L. Risbo, "On the design of tone free Σ-Δ modulators", to appear in *IEEE Transactions on Circuit and Systems — II*. The references are included in the bibliography section on page 173.

On the Design of Tone Free Σ-Δ Modulators

Lars Risbo

Abstract

Traditional one-bit sigma-delta modulators used for A/D and D/A conversion produce very predominant tones near half the sample rate which might intermodulate in the analog converter section and cause in-band tones. This paper demonstrates how the use of chaos can substitute dither as a means for extinguishing these tones. Especially, modulator feedback filters derived from noise transfer functions having an all-pass term seem very promising.

A.1 Introduction

Sigma-delta modulators are extensively used in the design of oversampling audio A/D and D/A data converters [1, 48].

One of the drawbacks of conventional sigma-delta modulators is the possible existence of tones in the output spectrum [40, 41, 57, 37]. A very common property is that when the modulator is fed with a dc-input, λ, the modulator spectrum contains a very strong out-of-band tone at the frequency $(1-\lambda)f_s/2$, where f_s is the sample rate [48, 40, 57]. Simulations have indicated that the tone close to $f_s/2$ causes the in-band noise power to vary periodically with the frequency λf_s (unpublished). Such periodic noise modulation is not observable on usual power spectra estimates but can be seen in the time domain and might be perceived as a tone by the human ear [40]. The frequency modulation of the tone near $f_s/2$ seems to be very general, i.e., sinusoidal input with frequency f produces characteristic side-band tones at the frequencies $f_s/2 - kf$ for integer k [48]. In addition, the high frequency tone and the side-bands can produce in-band intermodulation components produced by analog circuit nonlinearities [2, 24].

The conclusion from the observations mentioned above seems to be that precautions against the out-of-band tones must be taken if sigma-delta converters free from audible tonality are to be designed.

Enhanced suppression of both in-band noise and in-band tones can generally be achieved by increasing the order of the modulator, since this normally allows a higher attainable in-band loop gain. However, this has very little or no effect on the amplitude of the tones near $f_s/2$ since the out-of-band loop gain must remain nearly the same in order to preserve loop stability. Even eighth order modulators seem to suffer from these high frequency tones [48].

A different approach to tone suppression is the introduction of a degree of 'uncertainty' into the modulation. This is accomplished either by adding dither noise into the modulator loop [40, 41] or by the use of a chaotic modulator [41, 57, 37]. The purpose of this letter is to compare these two strategies with respect to the suppression of the tones near $f_s/2$. The advantage of a chaotic modulator is that no dither noise source is needed, thus eliminating the non-trivial task of designing a temperature independent noise source with a prescribed probability density function in analog implementations.

A.2 Feedback Filter Design from NTF-prototypes

Figure A.1 shows a typical sigma-delta modulator composed of a feedback filter (i.e, loop filter) $H(z)$ and a one-bit quantizer (i.e., signum function).

When the one-bit quantizer is crudely modeled as a unity gain with an additive noise source, the noise transfer function, NTF(z), from the noise source to the modulator output is given by:

$$\mathrm{NTF}(z) = \frac{1}{1 + H(z)} \tag{A.1}$$

The NTF describes how the quantization noise is spectrally shaped and suppressed in the base-band.

Sigma-delta modulators are typically designed by specification of a desired high-pass noise transfer function, i.e., an NTF-prototype [1, 49, 51, 58]. From an NTF-prototype given as a rational transfer function $\mathrm{NTF}(z) = A(z)/B(z)$ the feedback filter $H(z)$ can be derived using (A.1):

$$H(z) = \frac{1}{\mathrm{NTF}(z)} - 1 = \frac{B(z) - A(z)}{A(z)} \tag{A.2}$$

Since delay free loops around a quantizer are not implementable, $H(z)$ must have at least a one sample delay, i.e., the associated impulse response must have the property $h(k) = 0, \; k < 1$ [1, 41]. Equation (A.2) shows that this is achieved when the NTF-prototype is scaled such that $A(z)$ and $B(z)$ have the same highest order z-term. This necessary scaling causes a unity pass-band gain NTF-prototype to have a higher pass-band gain and a lower stop-band attenuation after scaling.

A good in-band noise suppression can be obtained by designing the NTF-prototype as a Chebychev II high-pass filter with unit circle zeros distributed in the base-band [1] (see also [58] for optimum unit circle zero locations). Such scaled prototype filters can easily be designed using the MATLAB Signal Processing Toolbox commands [49, 35]:

```
[A,B]=cheby2(N,Rs,fb,'high');
A=A/A(1)
```

where N is the filter order, Rs is the desired base-band attenuation in dB (prior to scaling) and fb is the upper base-band edge normalized with respect to $f_s/2$.

Increasing Rs will obviously improve the suppression of in-band noise and thereby the SNR; however, the necessary scaling causes the high frequency gain of the NTF-prototype to increase and, unfortunately, too much high-frequency noise destabilizes the modulator loop, i.e., the stable amplitude range of the modulator is reduced [1, 51, 58]. This phenomenon is caused by an overload of the simple one-bit quantizer, i.e., the quantizer output power is always unity and must be shared between the in-band signal and the circulating noise.

The feedback filter design is thus a trade-off between in-band noise suppression performance and loop stability. See [51, 58] for a discussion of design trade-offs and stability criteria for high-order modulators.

A.3 Chaotic Sigma-Delta Modulators

An Nth order sigma-delta modulator is a non-linear discrete-time dynamical system, which can be analyzed by observing trajectories in a corresponding N-dimensional state-space.

Generally, the state-space has a large number of periodic trajectories (i.e., limit cycles) with different periods [25].

It can be shown that if all poles of the feedback filter $H(z)$ are inside the unit circle, the limit cycles have attracting regions in the state-space which cause the modulator to lock asymptotically into strictly periodic modes [25].

If one or more of the $H(z)$ poles are outside the unit circle, the limit cycles will be unstable, i.e., even small perturbations in the state space will be amplified exponentially in time [57]. The sensitivity to initial conditions and a high density of limit cycles are important characteristics of chaos. Chaotic systems are unpredictable on long time scales and generate non-periodic outputs. Despite these properties, chaos does generally not guarantee that the quantization noise of sigma-delta modulators is free from tones. [57, 37].

The moduli of the poles outside the unit circle is a measure of the 'amount of chaos' in the modulator, i.e., how fast nearby points in state-space diverge per time step. The base two logarithms of the pole moduli are in fact equal to the so-called Lyapunov exponents used in chaos theory [25] and these exponents express the loss of state-space information in bits per time step.

A.4 NTF-prototypes with All-pass Terms

In order to obtain chaos, $H(z)$ must have at least one pole outside the unit circle. This is achieved when an NTF-prototype with zeros outside the unit circle is used, cf. (A.2).

An initial approach was to use a standard sigma-delta modulator and scale up the $H(z)$ unit circle poles with a certain factor; however this procedure reduces the in-band noise suppression significantly due to the resulting lower loop gain [57, 37].

Another way to obtain chaos is to use a minimum phase NTF-prototype and reflect one or more zeros to their reciprocal locations outside the unit circle, as demonstrated for FIR filters in [41]. The necessary scaling of the zero-reflected prototype scales up the magnitude characteristic by the product of the moduli of the reflected zeros, and this reduces the loop stability considerably. The stability was secured by adding more quantizer levels in [41]. The benefit of zero-reflection is that the in-band noise suppression is only reduced proportional to the necessary scaling.

An alternative to zero-reflection is to introduce an all-pass term into a standard IIR NTF-prototype with all zeros on the unit circle. A first order all-pass term must naturally have a (real valued) zero outside the unit circle and a pole inside the unit circle which is the reciprocal of the zero [44]. The advantages of an all-pass extended NTF-prototype are that the zeros are preserved on the unit circle for optimum in-band noise suppression and that the degree of chaos can be adjusted independently of the NTF magnitude characteristic by adjusting the all-pass zero location. A normalized all-pass term with a zero at $z = \alpha,\ |\alpha| > 1$ has the form:

$$\mathrm{NTF}_{ap}(z) = \frac{1 - \alpha z^{-1}}{1 - \dfrac{\alpha}{|\alpha|^2} z^{-1}} \tag{A.3}$$

This all-pass term may be multiplied on any normalized NTF-prototype thus giving a new normalized NTF-prototype with a filter order increased by one. Since the magnitude response of (A.3) is equal to $|\alpha|$, the loop stability deteriorates as $|\alpha|$ is increased. Consequently, 'more chaos' and thereby better tone suppression can be obtained at the expense of a lower stable amplitude range.

A.5 Design and Simulation Examples

In this section, three methods for suppressing the tones near $f_s/2$ are compared , i.e., NTF-prototype zero scaling, one-bit quantizer dithering, and finally, the use of an all-pass extended NTF-prototype.

For the sake of comparisons, the simulation examples will all be for sixth order modulators based on Chebychev II prototypes designed with MATLAB. All modulators are intended for 64 times oversampling, i.e., fb=$1/64 \cdot 20/22.05$ (suitable for e.g., 20 kHz base-band and $64 \cdot 44.1$ kHz sample rate). An additional design criteria was to ensure a stable input range up to 0.35 relative to full scale. The simulations use a DC-input of 1/256 corresponding to possible high frequency tones that might intermodulate and generate base-band tones at $f = 1/128 \cdot f_s/2$. The spectra shown in this section are obtained by averaging Kaiser-Bessel windowed 8k FFT power spectra for sequences of two mill. one-bit samples. The heavy averaging allows tones to be distinguished from random noise. In this context, the designation 'tone free' is used when possible tones are comparable to or below the noise floor provided by the given spectral resolution. An enhanced spectral resolution (e.g., larger FFT size) can naturally reveal more tones because the noise floor is lowered. For each spectrum the in-band noise power is computed by summing up the power of the appropriate FFT-bins.

The first modulator tested had a prototype designed with the parameters N = 6 and Rs = 120 dB. This modulator is stable for input amplitudes up to approx. 0.35 relative to full scale. Figure A.2 shows the simulated output power spectrum. The in-band noise power is approx. -142 dB (relative to full scale) corresponding to more than 21 bits of resolution. However, the tone seen near $f_s/2$ is no more than 6.7 dB below full scale — the tone is in fact stronger than any possible modulator input!

Experiments with scaling of the NTF-prototype zeros were then carried out. This resulted in a design based on a prototype with N = 6 and Rs = 90 dB. A zero scaling by a factor of 1.06 was necessary to suppress the high frequency tones below the noise floor. The resulting tone free spectrum is shown on Figure A.3. The in-band noise power is as high as -61 dB and this is a consequence of the NTF-zeros being removed from the unit circle. Furthermore, the experiments showed that this modulator is too unstable for any practical purposes.

The next approach was to add a dither noise source to the quantizer input. Obviously, this introduces more noise into the loop which consequently becomes more unstable. The NTF-prototype must therefore be designed with a somewhat lower Rs such that the same stable input range is preserved when dither is applied. The combination of N = 6, Rs = 95 dB and a spectral white dither source with a uniform amplitude density from -0.5 to 0.5 resulted in a tone free design with a stable input range up to approx. 0.35. The spectrum is showed on Figure A.4 and the in-band noise power is approx. -115 dB. A peak at -26.4 dB is seen at $(1 - 1/256)f_s/2$. When a lower dither span is applied, both a higher SNR and stable input range is obtained; however, the high frequency tone becomes stronger.

Finally, a sixth order modulator was obtained from a fifth order NTF-prototype extended by a first order all-pass term cf. (A.3). Experiments have shown that negative values of α are far more efficient than positive values in respect to suppression of high frequency tones — probably because an NTF-zero near $z = -1$ corresponds to $f = f_s/2$. The combination of N = 5, Rs = 80 dB and $\alpha = -1.25$ resulted in a tone free modulator with a stable input range up to approx. 0.35. The spectrum is showed on Figure A.5 and the in-band noise power is approx. -105 dB. A peak at -37.9 dB is seen at $(1-1/256)f_s/2$.

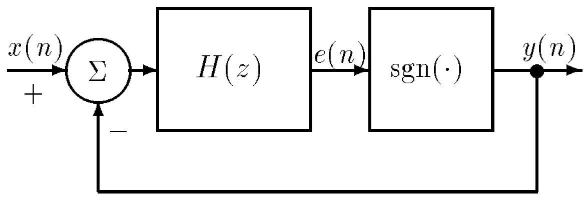

Figure A.1: Sigma-Delta modulator with feedback filter $H(z)$

The modulator has obviously a fifth order noise shaping amplitude characteristic and this is limiting the SNR. If the total order is increased to 7, it is possible to obtain a tone free design with a better SNR and the same stable amplitude range: the parameters N = 6, Rs = 92 dB and $\alpha = -1.2$ yields an in-band noise power of approx. -114 dB, i.e. almost the same performance as for the dithered sixth order example.

A.6 Conclusions

It has been demonstrated that the use of chaos effectively can substitute dither as a means for suppression of the predominant tones near half the sample rate in one-bit sigma-delta modulators. In particular, the use of a general all-pass term seems to be very advantageous. Investigations have shown that a first order all-pass term with a real valued negative zero outside the unit circle is a feasible choice. The price paid for the tone suppression is that both dithering and the use of chaos reduce the loop stability significantly. In order to restore stability, the in-band loop gain must be reduced resulting in a lower attainable SNR for a given modulator order. However, this can to some extend be remedied by increasing the modulator order.

The presented examples indicate that the dithered modulator obtains a 9 dB better SNR than a chaotic modulator for the same total order and stable amplitude range. However, this is to some extend counterbalanced by the high-frequency tone being approx. 11 dB weaker for the chaotic modulator example than for the dithered. Future research should therefore attempt to uncover the exact relations between tone strength, dither span and the all-pass zero modulus.

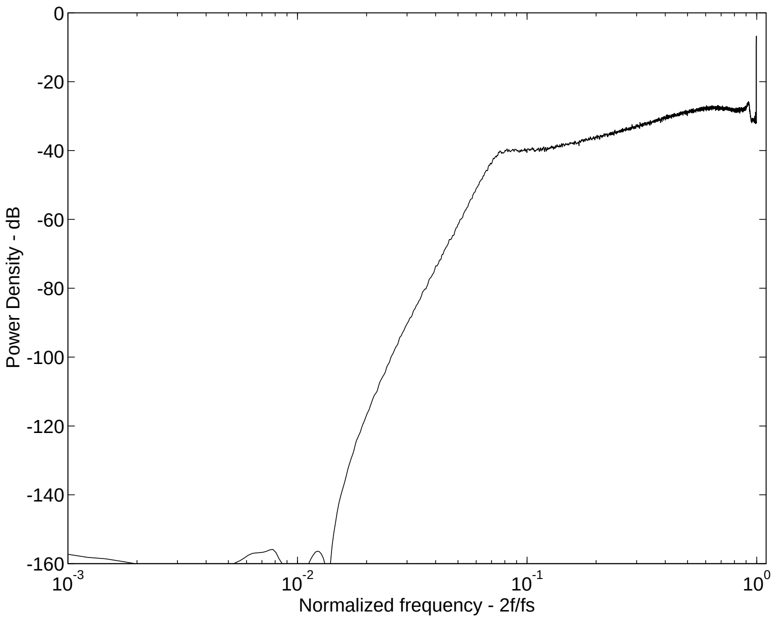

Figure A.2: Power spectrum from a simulation of a sixth order modulator with Rs = 120 dB.

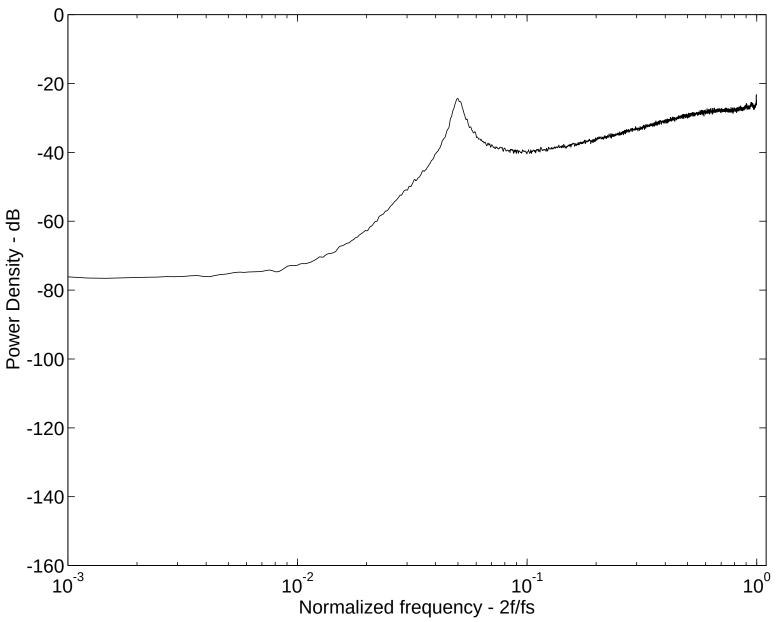

Figure A.3: Power spectrum from a simulation of a sixth order modulator with Rs = 90 dB and an NTF-zero scaling of 1.06.

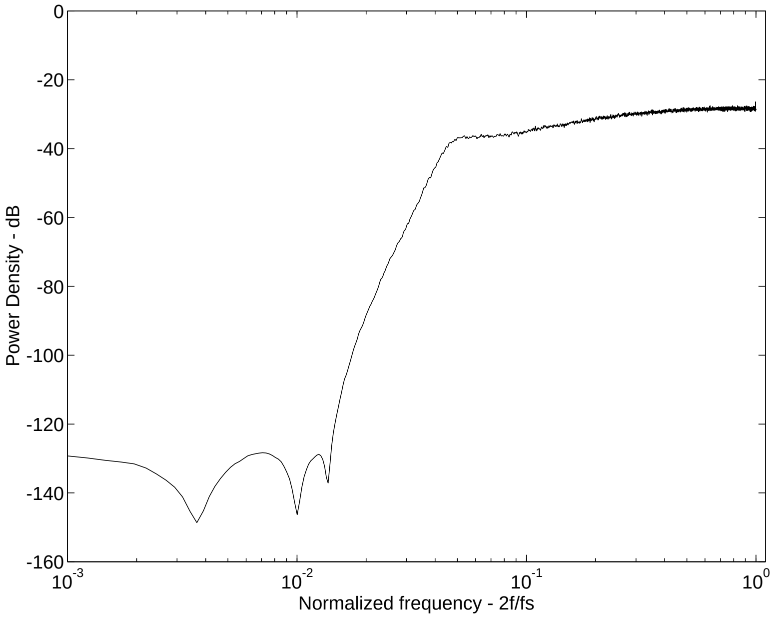

Figure A.4: Power spectrum from a simulation of a sixth order modulator with Rs = 95 dB using dither uniformly distributed over the interval [-0.5,0.5].

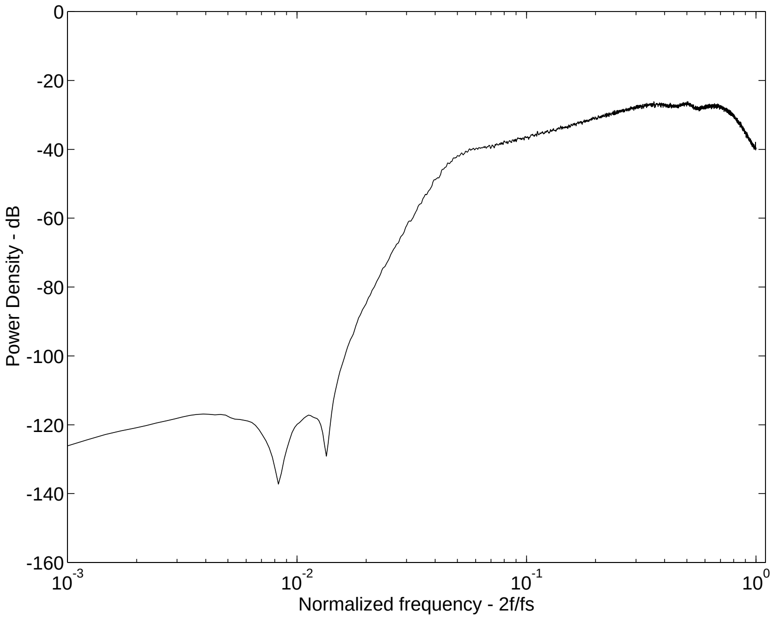

Figure A.5: Power spectrum from a simulation of a fifth order modulator with Rs = 80 dB extended to sixth order using an all-pass term with $\alpha = -1.25$.

Appendix B

Appendix for Chapter 9

This appendix is connected to chapter 9 and is a reprint of reference [52]: L. Risbo, "Improved stability and performance from Σ-Δ modulators using one-bit vector quantization," *IEEE Proc. ISCAS'93*, pp. 1361–1364, May 1993.

Appendix C

Implementation of an Audio DAC

This appendix is a reprint of the reference [48]: L. Risbo, “FPGA based 32 times oversampling 8th order sigma-delta audio DAC” presented at the 96th Audio Engineering Society Convention, Amsterdam, The Netherlands, 1994 February 26 — March 01, Preprint 3808.

The paper presents a Σ-Δ D/A-converter which was developed during the Ph.D. study. The references are included in the bibliography section on page 173.

Abstract

A novel Sigma-Delta based audio DAC system is presented. The converter operates at only 32 times oversampling (1.4 MHz) and uses an 8th order Sigma-Delta modulator implemented in a single FPGA circuit. Two different DAC circuits have been implemented: One using switched capacitor charge pulses and one using non-return-to-zero pulses. A VCXO based PLL ensures sampling clock recovery free from deterministic jitter components.

C.1 Introduction

In recent years, Sigma-Delta based oversampling A/D and D/A converters have become very popular for demanding audio applications [38, 1, 61, 23]. Especially, converters using two level signal encoding, i.e., one-bit single loop Sigma-Delta converters, have attractive properties: The almost perfect level linearity and the lack of differential non-linearity. Furthermore, these devices do not require high precision trimming of current sources as for the conventional multibit types.

A Sigma-Delta modulator (SDM) consists of a one-bit quantizer (i.e., a comparator) and a suitable loop filter – or feedback filter. The one-bit output signal of the quantizer is forced to reproduce the input signal due to the feedback loop. The resulting output signal will consequently consist of the modulator input signal and a quantization noise component which is spectrally shaped by the feedback loop. The use of heavy oversampling enables the quantization noise to be removed from the audio band, and therefore concentrated at high frequencies. When the resulting one-bit signal is low-pass filtered, the input signal is reproduced with a high signal-to-noise ratio (SNR).

A Sigma-Delta D/A-converter uses a digitally implemented Sigma-Delta modulator to resample the high resolution input signal (i.e., 16–20 bit) and produce a one-bit signal at a sample rate typically 64–1024 times higher than for the input. The one-bit signal is subsequently fed to a one-bit DAC, i.e., a switch, and the resulting two-level current or voltage is low-pass filtered in the analog domain.

Early Sigma-Delta based DACs used a low order feedback filter and a high oversampling ratio (e.g., second order modulator with 256 times oversampling [38]). A consequence is that the design of the analog converter section becomes very demanding due to the very high sample rate (> 10 MHz). Furthermore, the quantization noise of second order SDMs is generally not very random and is somewhat correlated with the input signal. However, this can be remedied by proper dithering.

The current trend is to use high order feedback filters. Recently, a 64 times oversampling 5th order DAC has been reported [61]. High order feedback filters allow a more frequency selective noise-shaping and consequently a lower oversampling ratio is needed. In addition, the resulting quantization noise becomes less correlated with the audio signal due to the higher complexity in the one-bit encoding process. These converters have properties which are comparable to those of analog amplifiers: The noise is almost independent of the input and is therefore not perceived as distortion of the input signal. Since a Sigma-Delta modulator of course is a deterministic system, the quantization noise is not random; however, high-order feedback makes the relationship between input signal and the noise inscrutable to the ear. Conventional multibit converters introduce, unlike the one-bit types, an error which is directly dependant on the signal amplitude or digital input code.

The objective of this paper is to present a new converter based on an 8th order modulator operating at only 32 times oversampling, i.e, approx. 1.4 Mhz for a usual CD-standard input. The entire converter, including standard SP/DIF digital audio interface and analog post filtering, is exclusively made of standard off-the-shelf components. The digital modulator section is implemented in a general purpose Field Programmable Gate Array (FPGA) circuit.

C.2 Converter Overview

The converter consists of six blocks shown on Figure C.1:

- Digital interface receiver
- Interpolation filter
- 8th order SDM
- One-bit DAC
- Analog post filter
- Clock jitter attenuator (VCXO PLL)

The digital interface receiver (Crystal CS8412) demodulates the standard SP/DIF or AES/EBU digital audio format and regenerates a clock and a serial 16-bit data signal. Subsequently, this signal is fed to an interpolation chip (NPC SM5803) which oversamples the signal eight times. The resulting 18-bit output signal is fed to the 8th order Sigma-Delta modulator, which oversamples the signal four times additionally and performs the one-bit encoding. The analog section of the converter consists of a one-bit DAC operating on the 32 times oversampled modulator output and a post filter which removes the out-of-band quantization noise. The last block is a voltage controlled crystal oscillator (VCXO) based Phase Locked Loop (PLL) for clock jitter attenuation. The PLL suppresses timing jitter on the clock recovered by the digital interface receiver which otherwise would cause a degradation of the DAC performance. In the following sections, the Sigma-Delta modulator and the analog section will be described in greater detail along with some background theory and design considerations.

C.3 Sigma Delta Modulator

C.3.1 Design and Analysis

The modulator topology used for the DAC is shown on Figure C.2 for even modulator order, N. This topology is known as the *multiple feedback* type [1, 55, 49]. The input signal is fed trough a cascade of delaying discrete time integrators to the one-bit quantizer which output is distributed back to each of the integrators in the chain using the weight factors b_i. Additional local feedback is applied around two adjacent integrators as indicated by the a_i-coefficients. This topology has a very regular structure which is a cascade of simple second order sections. Furthermore, this topology has very low sensitivity to coefficient errors [49].

The local feedback introduced in the integrator chain implies that each of the a_i-coefficients introduces the following complex open loop pole pairs:

$$d_i = 1 \pm j\sqrt{a_i} \tag{C.1}$$

Note that these poles are generally outside the unit circle implying that the integrator chain is totally unstable without the feedback. Furthermore, poles outside the unit circle will destabilize possible limit cycles making the modulator chaotic and non-periodic [49].

By linearized analysis, the modulator can be characterised by a Signal Transfer Function, $STF(z)$, and a Noise Transfer Function, $NTF(z)$ [1, 54]. The $STF(z)$ shows how the input signal is affected and $NTF(z)$ shows the noise-shaping characteristic for the quantization noise.

One advantage of the multiple feedback modulator is that $STF(z)$ becomes a low-pass filter [49]. This relaxes the requirements of the preceeding interpolation filter. Especially for a high modulator order (e.g., 8th order) the last stages of the interpolation filter can be replaced by a simple zero order hold circuit [49].

When a modulator is designed, the a_i coefficients are specified at first such that the open loop poles are distributed within the audio-band which ensures good noise suppression for $NTF(z)$. Subsequently, the b_i-coefficients are chosen for an acceptable trade off between SNR and loop stability. Furthermore, the choice of b_i must also ensure that the resulting $STF(z)$ is reasonably flat in the audio-band.

The implemented modulator uses coefficients obtained using tools described in [54, 51]. The 8th-order modulator was optimized for 32 times oversampling, a flat $STF(z)$ and stable operation up to output amplitudes of 0.4 relative to full scale. The a_i-coefficients used were rounded to power of two values allowing binary shift operations to be used instead of general multiplications.

The maximum stable amplitude range (MSAR) is a very important design parameter. Generally, any MSAR can be obtained by chosing the b_i-coefficients properly [49]. However, optimum SNR is obtained when MSAR is around $0.3 - 0.35$ relative to full scale [49]. If MSAR is lower, the SNR will deteriorate due to the lower maximum signal power. For a higher MSAR, the in-band noise power increases faster than the signal power, and consequently, the SNR decreases. Modulators with a high MSAR also exhibit higher intrinsic harmonic distortion and more variation of the in-band noise power when varying the signal power [49]. When other analog noise sources are taken into consideration, the optimum MSAR may be shifted towards higher values.

Figure C.3 shows power spectra estimates of the modulator output obtained by simulations using the actual 24-bit integrators used in the implemented modulator. Figure C.4 shows the in-band RMS noise as function of the output DC-amplitude. This graph is obtained from simulations of the modulator with an extremely slowly increasing ramp input. Notice the absence of spikes and irregularities on this curve. This type of plot is rather revealing for the modulators ability to decorrelate the noise from the input. Furthermore, MSAR can be found using these ramp input simulations. In this case MSAR was 0.412. The SNR for sinusoidal input with amplitude 0.4 is estimated to 97 dB using Figure C.3.

C.3.2 Modulator Implementation

Figure C.2 shows that for the multiple feedback type modulator topology, only one-bit values, i.e., -1 or 1 are multiplied on the b_i-coefficients and, consequently, only an ADD or a SUB operation is needed depending on the quantizer output. Furthermore, when the a_i-coefficients are designed as power of two values, binary shift operations can replace general multiplications. The absense of general multiply operations makes the multiple feedback modulator very well suited for implementation in programmable logic.

The 8th order modulator was implemented in a single 3000 gate equivalent Xilinx XC4003 FPGA circuit using hardware efficient bit-serial arithmetic. The implementation consists of four almost identical second order modulator sections, a shared control circuit and a signal input conditioner. A fully synchronous design was used operating on a $384f_{in}$ input clock. This gives 12 clock cycles for processing of each output sample when using 32 times oversampling.

An other FPGA implementation of a Sigma-Delta modulator has been reported in [27]. In [27] a fourth order multiple feedback modulator was implemented in a single 9000 gate equivalent Xilinx XC3090 FPGA using bit-parallel arithmetic. However, the local feedback introduced by the a_i-coefficients was not implemented.

The arithmetic operations are performed serially operating on 2-bits for each clock cycle allowing an integrator resolution of $2 \cdot 12 = 24$bits. Each integrator consists of a 12-bit deep and 2-bit wide shift register and a number of 2-bit adders. This provides a state-space of totally $8 \cdot 24 = 192$-bits which ensures extremely long and complex modulator idle pattern sequences. Binary scaling shifts are used between the integrators in order to fully utilize the dynamic range of the integrators. The binary (right) shift operations are implemented by taking intermediate outputs from the integrator shift registers. Sign extension of the right shifted data stream is implemented using a multiplexer which retransmits the integrator sign bit when the most significant bit is reached.

Figure C.5 shows a simplified diagram of one of the four second order sections. Each section performs five ADD/SUB-operations and three shift operations i.e., one for each of the integrators and a shift for the a_i-coefficient multiplication. This corresponds to a total computational capacity of approx. 50 mill. 24-bit operations per second for the entire modulator.

The control circuit generates a number of shared controls signals which are used to control the modulator. A modulo 12 counter genererates a 4-bit address sequence which is fed to a number of look-up ROM tables producing different control signals. For instance, the b_i-coefficients are supplied bit-serially to the integrators from eight ROM-tables. Since the used FPGA holds programming information in a SRAM memory, new modulator coefficients can be programmed in miliseconds. In fact, during the design and test phase of the project, the FPGA was programmed by downloading information from a PC. A PC-program was developed which translates loop filters designed in a high-level environment (MATLAB) into boolean equations for the FPGA. This enables many loop filters to be tested under realistic conditions without redesigning silicon.

The used FPGA is organized as a 10 by 10 array of Configurable Logic Blocks (CLBs) which each contains two D-type latches and two four input arbitrary logic function generators. Additionally, the chip contains a large number of input/output blocks (IOBs) which each contains an input and an output latch connected to a pin. The ADD/SUB blocks in Figure C.5 use each two CLBs and the MUX and ROM functions use each one CLB. A total of 15 CLBs are used for each second order section. In addition, $2 \cdot 2 \cdot 12 = 48$ latches are needed for each second order section. These latches are partly taken from the remaining CLBs and partly from the IOBs of unused pins.

The input conditioner accepts the 8 times oversampled 18-bit serial input signal from the interpolation filter and loads a parallel input register. In order to upsample to 32 times oversampling, the content of the input register is retransmitted serially 4 times to the modulator for each input sample.

No hardware precausions against modulator instability were taken in the implementation except that proper coefficient scaling ensures that the output amplitude is limited to 0.4 relative to full scale for maximum digital input. However, special precausions have been taken in the design of the loop filter [54] such that the modulator operates reliably up to the 0.4 amplitude limit. The modulator has successfully been tested extensively with both full scale sinusoidal input and music programme input.

C.4 Analog Section

The objective of the analog design was to achieve superior sound quality. Two different one-bit DACs and two different analog post filters have been designed using standard discrete components.

C.4.1 One-bit DAC overview

The one-bit DAC converts the binary bit-stream from the modulator into an analog waveform which can be interpreted as a convolution between the digital one-bit signal $p(k)$ and an analog pulse, $g(t)$:

$$d(t) = \sum_k p(k) g(t - k \cdot T) \tag{C.2}$$

where T is the sampling time interval and $p(k) \in \{-1, 1\}$.

In the frequency domain, the convolution with $g(t)$ will act as a filtering of the digital one-bit signal. Fundamentally, two one-bit DAC types can be discriminated, i.e., the pulse type and the hold type. Pulse type DACs generate concentrated current pulses by charging or discharging capacitors between reference voltage levels [38, 61]. The hold type DACs hold a reference voltage or current level in a certain period of time. Usually the hold time is derived from the system clock. The hold type DAC can further be divided into return-to-zero (RTZ) types [23] and non-return-to-zero (NRZ) type. The latter holds the reference level for precisely one sampling time interval, and RTZ types return to zero output before the next input sample. Consequently, all three types can be destinguished by their corresponding $g(t)$ pulses.

C.4.2 Spectral DAC Properties

The three types of DACs have a number of different properties. The most apparent differences are the spectral properties induced by the $g(t)$ waveforms. For the pulse type DAC, $g(t)$ approaches a delta function which has a very wide-band spectrum. Consequently, the periodic spectrum of the digital DAC-input found at multiples of the sampling frequency is preserved up to high frequencies in the analog DAC-output. This means that the succeeding analog post filter must be able to handle wide-band signal and noise components without generating intermodulation components folded back into the audible range [28].

The hold type DACs have generally more intrinsic low-pass filtering due to longer $g(t)$ pulses. The amplitude response for the Fouriertransform of a unity amplitude square pulse of duration τ is given by:

$$|G(f)| = \tau \frac{\sin \pi f \tau}{\pi f \tau} \tag{C.3}$$

In particular, for $\tau = T$, $G(f)$ has zeros for multiples of the sampling frequency $f_s = 1/T$. This means that NRZ DACs have a first order suppression of the image spectra at multiples of the sampling frequency. For 32 times oversampling, the image spectrum at f_s will be attenuated at least $37\,dB$. The author believes that this image spectrum attenuation is very beneficial for the sound quality because much high-frequency distortion is avoided in the the succeeding analog signal processing. In fact, the usual multibit type converters use NRZ conversion; hence, they have the same attenuation of image spectra. It is very likely that the improvement in sound quality obtained when oversampling multibit DACs were introduced is attributable to the improved image spectra attenuation which is a consequence of oversampling.

For a RTZ DAC, τ will be less than T and the zeros of $G(f)$ will generally not be at multiples of f_s. Furthermore, the output signal amplitude scales directly with the hold time τ, and consequently, the analog SNR will deteriorate when other analog noise sources are present.

C.4.3 Clock Jitter Sensitivity

When clock timing jitter is taken into consideration, the pulse type and hold type DACs will show different properties [14]. Clock jitter is the designation for timing errors of the sampling instants which cause errors in the reproduced analog waveforms. It will be assumed that the sampling instants are displaced in time by the jitter sequence $t(k)$, i.e., the sampling instants are: $s(k) = k \cdot T + t(k)$. For the NRZ DAC, the hold time will be modulated by the jitter sequence, since the output signal only changes state at the sampling instants. It can be shown [14] that the error induced by the jitter sequence can be modeled as the following digital error sequence:

$$e_{\mathrm{NRZ}}(k) = (p(k) - p(k-1))\frac{t(k)}{T} \tag{C.4}$$

This equation shows that the error is proportional to the modulator output frequency due to the differencing operation. Notice that the equation also applies for usual multibit converters. The problem when dealing with one-bit DACs is the extreme content of high-frequency quantization noise in the $p(k)$-signal which has more noise power than signal power. When the jitter sequence is a wide-band and uncorrelated noise, the error signal of (C.4) will be almost white. The variance of the error sequence is then approximately the product of the jitter variance and the variance of $p(k) - p(k-1)$. Simulations have shown that the variance of the difference signal $p(k) - p(k-1)$ is approx. 2.7 for the used 8th order modulator with zero input. The in-band SNR for R-times oversampled sinusoidal output with amplitude A and RMS jitter t_{RMS} is consequently:

$$\mathrm{SNR}_{\mathrm{NRZ}} = 10 \log \frac{R\ A^2/2}{2.7\left(\frac{t_{\mathrm{RMS}}}{T}\right)^2}\ \mathrm{dB} \tag{C.5}$$

In order to obtain a SNR of e.g., 100 dB with 32 times oversampling ($f_s = 1.4\,\mathrm{MHz}$) and sinusoidal output with amplitude 0.4, the RMS wide-band jitter must be less than 7 picoseconds !

The pulse type DACs have a different sensitivity to clock jitter due to the fact that the pulse waveform is typically the result of a capacitor charge or discharge which is independent of the sampling time instants.

In [14] it has been shown that the equivalent digital jitter error signal for pulse type DACs could be interpreted as the product signal $p(k)t(k)$ filtered by a differentiator with transfer function $H(f) = j2\pi f$. This means that the jitter error is noise-shaped and suppressed significantly for low frequencies. Consequently, pulse type DACs can tolerate much more wide-band clock jitter.

The in-band jitter noise power can be calculated by integrating the shaped noise power density in the audio band. When assuming that the error generating sequence $p(k)t(k)$ is white with variance t_{RMS}^2, one arrives at the following SNR figure:

$$\mathrm{SNR}_{\mathrm{pulse}} = 10 \log \frac{3\ R^3\ A^2/2}{\pi^2 \left(\frac{t_{\mathrm{RMS}}}{T}\right)^2}\ \mathrm{dB} \tag{C.6}$$

In order to obtain a SNR of e.g., 100 dB with 32 times oversampling ($f_s = 1.4$ MHz) and sinusoidal output with amplitude 0.4, the RMS wide-band jitter must be less than 200 picoseconds. This shows that the pulse type DACs are considerably less sensitive to wide-band clock jitter. However, it should be emphasized that the shown calculations only have taken the interaction between the quantization noise and the jitter into account. It has been pointed out in [14] that pulse type DACs can be more sensitive to deterministic jitter than NRZ DACs for low frequency signal output.

C.4.4 Sensitivity to Unequal Rise and Fall Times

A major drawback of NRZ one-bit DACs is their high sensitivity to unequal rise and fall times of the output signal. The problem is that the $g(t)$ pulses will depend slightly on the previous output code in case of unsymmetrical transition times. This phenomenon gives rise to noise and distortion of the analog output.

The resulting error could be interpreted as a kind of 'auto jitter', i.e, the resulting transition time or sampling instant depends on the DAC input code. If the rise time is e.g., longer than the fall time, the equivalent transition time will be delayed for a positive input code. This corresponds to a jitter sequence of $t(k) = \alpha T(p(k)+1)/2$ and this results in an error sequence of $\alpha(p(k) - p(k-1))(p(k)+1)/2$ cf. (C.4). This signal is equal to 2α when $p(k)$ changes from low-to-high and zero otherwise. The error signal contains obviously second harmonic distortion due to the squaring of the signal $p(k)$. Furthermore, the error signal contains wide band noise due to intermodulation distortion of the high-frequency noise in $p(k)$. The error signal spectrum can be asessed by simulations. It is estimated by simulations that the rise and fall time difference must be less than 24 picoseconds for 100 dB SNR with the actual modulator. The error can typically be reduced considerably by a differential design giving symmetric transitions.

Pulse type and RTZ type DACs do not have this intrinsic error mechanism as far as the DACs are able to settle completely between the output pulses; however, incomplete and unsymmetric settling might cause a similar error.

C.4.5 Deterministic Components Near $f_s/2$

Almost all Sigma-Delta modulators produce variuos tones near $f_s/2$. If the modulator is fed with a constant giving a mean output value of m_p, a spectral peak will be present in the output spectrum at the frequencies $(1 \pm m_p)f_s/2$ (see Figure C.3) [49, 24]. The peaks might intermodulate in the analog circuitry and cause audio-band tones at $m_p f_s$.

This mechanism is a kind of frequency modulation. Consequently, if the modulator input is sinusoidal with frequency f, the modulator output should contain peaks or sidebands at $f_s/2 \pm nf$ for integer n. This conjecture is confirmed by Figure C.6 which shows an output spectrum for the used modulator with sinusoidal input with frequency $0.025 f_s/2$ and amplitude 0.2 relative to full scale. Again, nonlinearities or clock jitter could turn these peaks into harmonic distortion in the audible range. Obviously, interfering signals with frequency $f_s/2$ should be avoided as stated in [24].

In [54] it is revealed how Sigma-Delta modulators generally can be designed such that thay do not produce spectral peaks near $f_s/2$.

C.4.6 Pulse DAC Implementation

A pulse type DAC was implemented using discrete transistors and CMOS switches. Figure C.7 shows a simplified schematic of the pulse DAC. The DAC operates with only one capacitor which is either charged to a positive or a negative reference level depending on the digital one-bit input code. In between each sample, the capacitor is discharged to ground. The capacitor charge/discharge and reset operations are controlled by two non overlapping clock phases. The current pulses used for the capacitor charge or discharge are taken directly from the DAC output using bipolar transistors coupled as current conveyers. These transistors are furthermore isolated from the DAC output by two cascode transistors. This ensures that the DAC output signal cannot modulate the reference levels used for the switched capacitor. The entire DAC is thus implemented in current mode topology, and furthermore, the design is totally based on feedforward coupling. This ensures a high bandwidth and a good high frequency linearity.

C.4.7 Current Switch NRZ DAC Implementation

A NRZ current mode DAC was implemented using discrete bipolar transistors. A simpified schematic is found in Figure C.8. The heart of the DAC is a current switch based on a diode bridge. The incoming TTL-signal from the CMOS latch can either draw current from an upper or a lower reference current cource depending on the logical state. The current source which is not feeding the latch-output must feed the input transistors Q1 and Q2 which transmit the current to the output node through folded cascodes (Q3 and Q4). This arrangement ensures that the DAC output voltage does not modulate the reference currents. Again, the circuit is free from feedback and operates primarily in current mode.

The threshold level for the current switch can be adjusted by varying the base voltages of the input transistors Q1 and Q2. This feature makes it possible to adjust the delays of the low to high and high to low transitions relative to the clock signal. Consequently, this trimming option can null out the noise and distortion due to unequal rise and fall times. The adjustment can be done manually simply by means of ear-phones. However, the adjustment seems to be slightly temperature dependent.

C.4.8 Analog Post Filter

The analog post filter 'decodes' the one-bit signal, i.e., removes the out of band quantization noise components which could cause distortion in the succeeding analog circuitry. The post filter should be able to handle noise and deterministic components far into the MHz range without generating intermodulation components. These requirements call either for a passive design or the use of very high bandwidth and non slew-rate limiting active devices.

The use of 8th order noise-shaping makes it very difficult to suppress the quantization noise which rises very abruptly just over the audible range. This is the major drawback of low oversampling and high-order DACs. The filter characteristic used is a 5th-order Butterworth filter with a cut-off frequency at 25 kHz. Additionally, a transmission zero at approx. 58 kHz has been introduced in order to improve the noise suppression. Figure C.9 shows the amplitude response. The filter has approx. 1.5 dB amplitude loss at 20 kHz. The phase response can be corrected in the digital domain due to the minimum phase characteristic.

The filter results in a wide-band SNR of only 48 dB due to residual noise between 20 and 70 kHz which is difficult to suppress. This residual out-of-band noise is considered as harmless to the succeeding amplifiers and speakers, though a better suppression would be desirable.

The post filter has been implemented both as a passive LCR design and as an active design (Figure C.10). The passive design uses inductors with large air-gapped ferrit cores and high voltage polypropylene film capacitors for best linearity. The active design consists of a third order all-pole Sallen-Key filter followed by a Sallen-Key twin-T filter giving the remaining two complex poles and zeros. The active elements are based on AD811 current feedback op-amps with 140 MHz bandwidth, 2500 V/μS slew-rate and 100mA output current capability. The impedance level for both the active and passive filter is 300 Ω. Notice, the input capacitors for both filters which limit the DAC slew-rate.

C.4.9 Jitter Attenuator

A narrow band Phase Locked Loop (PLL) has been implemented which suppresses timing jitter on the recovered clock. It has been shown [14, 36] that the clock recovered from a SP/DIF interface can contain large amounts of deterministic jitter components due to both poor integrated circuit decoupling and insufficient interface bandwidth. It is strongly suspected that these jitter components can deteriorate the sound quality of a DAC.

The PLL is based on a Voltage Controlled Crystal Oscilator (VCXO) operating at $384f_{in}$. The used phase detector (AD9901) and PLL loop filter gives a closed loop cut-off frequency at 20 Hz. Jitter is asymptotically attenuated by 18 dB/Octave above approx. 100 Hz. The purified clock is fed directly to the one-bit DAC latches.

C.5 Measurement Results

An Audio Precision System One measurement system was used to meausure the converter performance. The test system supplies a digital test signal (sinewave) to the SP/DIF input of the converter and the resulting analog output is analysed. Figure C.11 shows the output spectra for a 0 dB full scale 1 kHz dithered 18-bit input for the pulse DAC with the active post filter and the NRZ DAC with the passive post filter. Notice the low harmonic distortion of the NRZ DAC with passive filtering. With -60 dB input, the signal to noise ratio below 20 kHz was measured to 34.1 dB and 31.4 dB, for the pulse and NRZ DAC, respectively. This corresponds to dynamic ranges of 94.1 dB and 91.4 dB. The figures include quantization noise present in the dithered digital input.

Figure C.12 shows the level linearity of the pulse DAC and active filter.

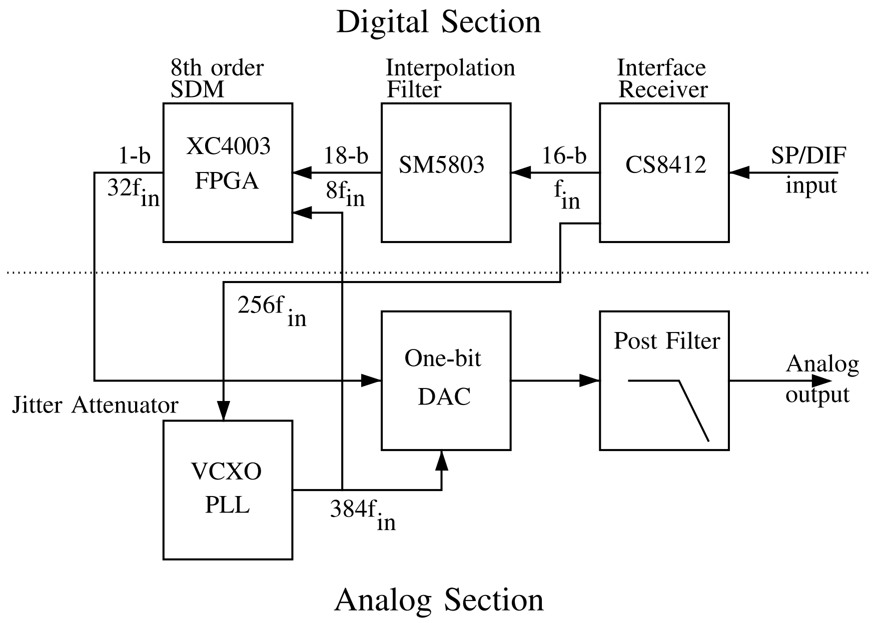

Figure C.1: D/A converter overview showing three digital and three analog building blocks.

C.6 Conclusions

The implementation of an entire 8th order and 32 times oversampling one-bit Sigma-Delta audio DAC system has been addressed. The advantage of high-order encoding is better quantization noise decorrelation, and the advantage of a low oversampling ratio is a better linearity in the analog section. The disadvantage of this approach is the demand for a high-order analog post filter in order to suppress the sharply rising quantization noise. However, digital phase correction can be implemented.

Furthermore, it has been demonstrated that the digital modulator can be implemented in a single general purpose FPGA circuit using bit-serial arithmetic.

The sound quality of this DAC system has been judged subjectively as being superior to a number of tested commercial designs including both multibit and one-bit types. Especially, the sound from the NRZ DAC with the passive filter is very transparant. The author belives that this is attributable to both the choice of modulator and the careful and simple feedforward analog design , including the use of jitter attenuation and passive filters.

Acknowledgements

The author is very grateful to Thomas Mørch who inspired the author to use the diode bridge current switch and designed the printed circuit boards for the converter.

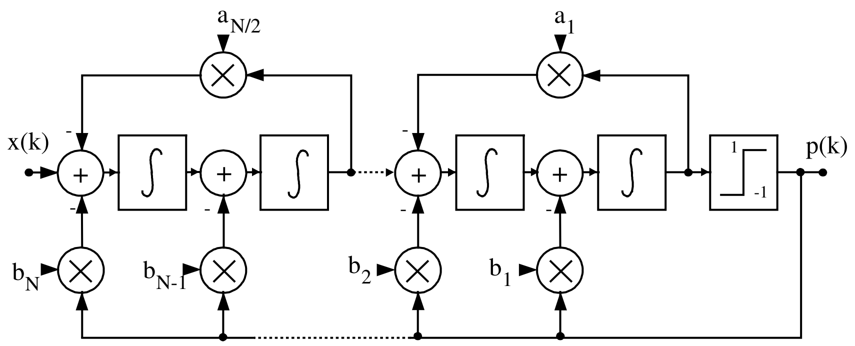

Figure C.2: Sigma-Delta modulator with multiple feedback topology for even order N. $x(k)$ is the input and $p(k)$ is the output. The blocks marked $\int$ are time-discrete delaying integrators with transfer function $1/(z+1)$.

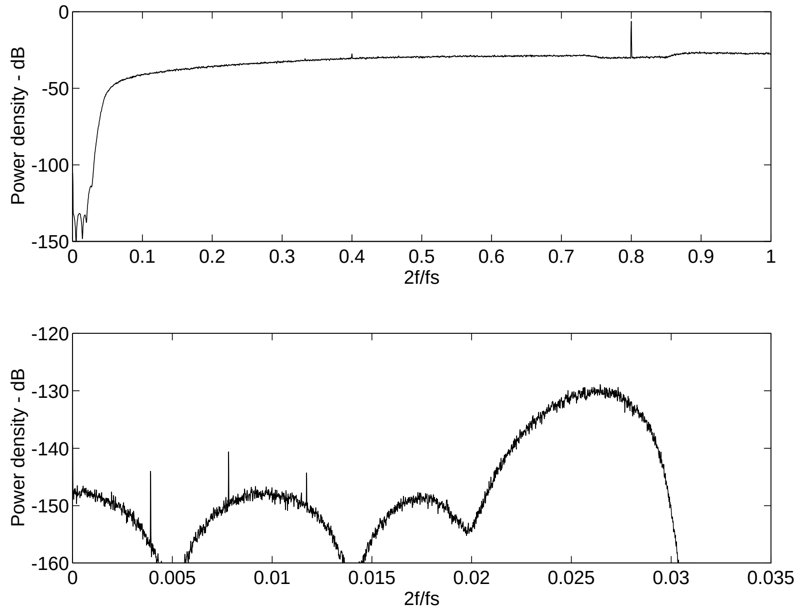

Figure C.3: Power density spectra estimated using averaged 8k FFTs with Kaiser-Bessel window on bit-exact 24b simulations of the modulator output (0 dB corresponds to a full-scale sinewave). Upper graph: Full spectrum of modulator with DC-input 0.2 (2 mill. samples). Lower graph: 32 times decimated modulator output with DC-input 1/512 (500.000 decimated samples).

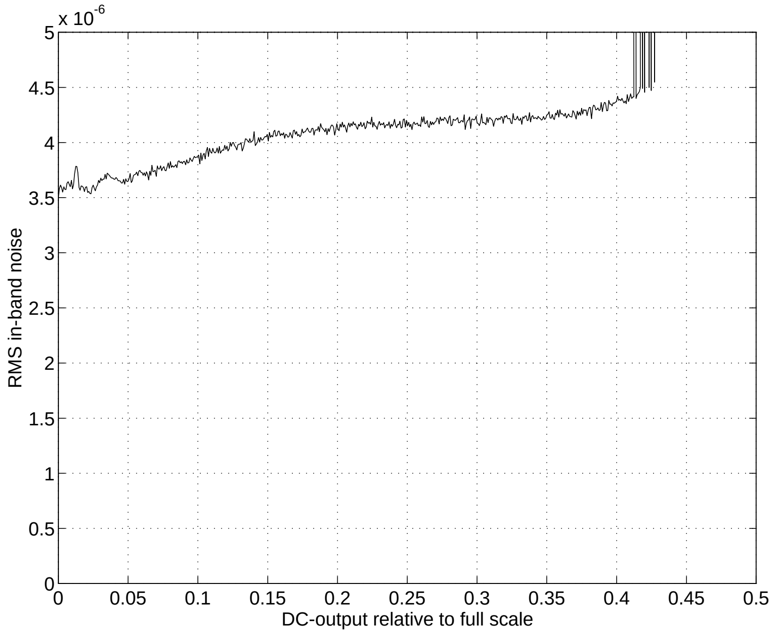

Figure C.4: In-band modulator output RMS noise versus DC-output measured by bit-exact 24b simulations and decimation. Each of the approx. 512 data points is based on decimation of 1.6 mill. modulator samples. The spikes at the end of the plot are the onset of instability.

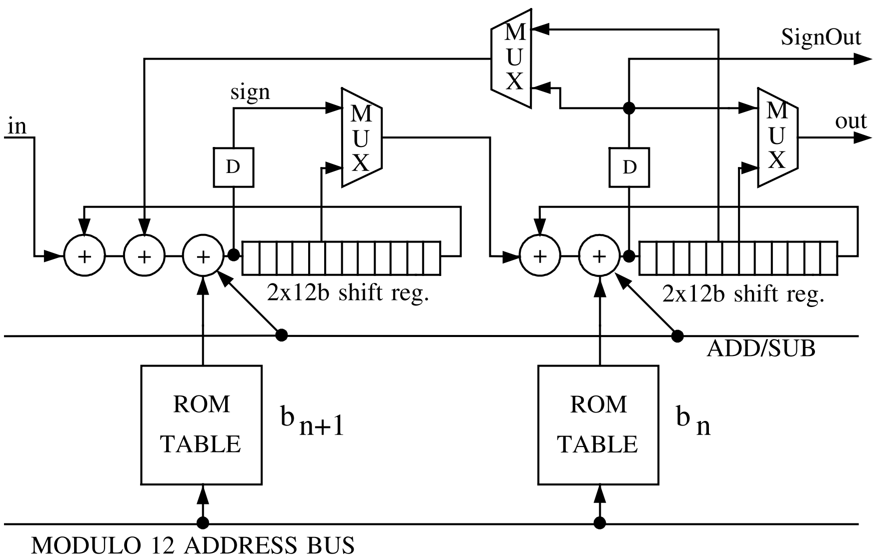

Figure C.5: FPGA implementation of a second order multiple feedback modulator section. Four of these sections are cascaded to produce an 8th order modulator. The 'SignOut' signal of the last section is the output, $p(k)$, which is fed back to the ADD/SUB control line.

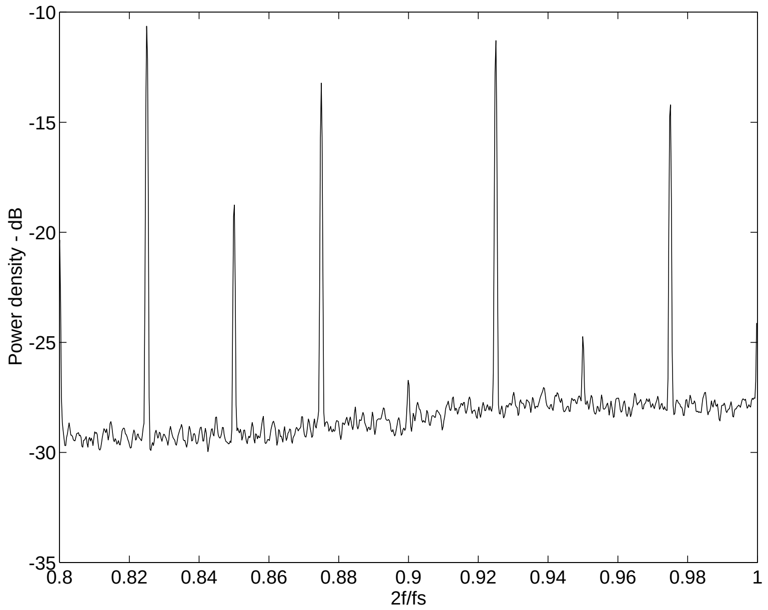

Figure C.6: Modulator power spectrum for sinusoidal input with frequency $0.025 f_s/2$ and amplitude 0.2. Notice the side-bands at $2f/f_s = 0.975,\ 0.95,\ 0.925$ etc.

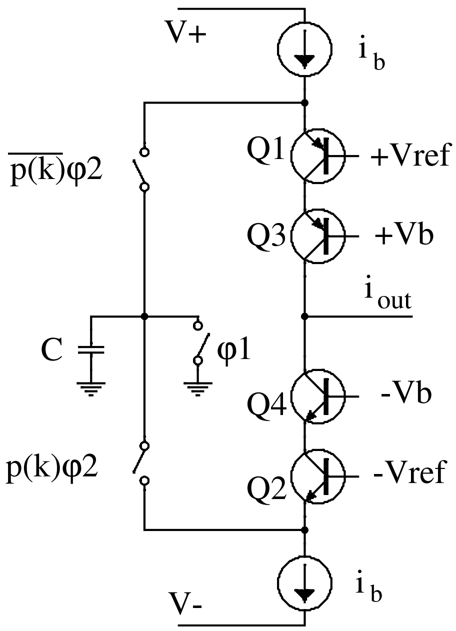

Figure C.7: The implemented pulse type one-bit DAC. The circuit is controlled by the modulator output $p(k)$ and two non-overlapping clocks, φ_1 and φ_2.

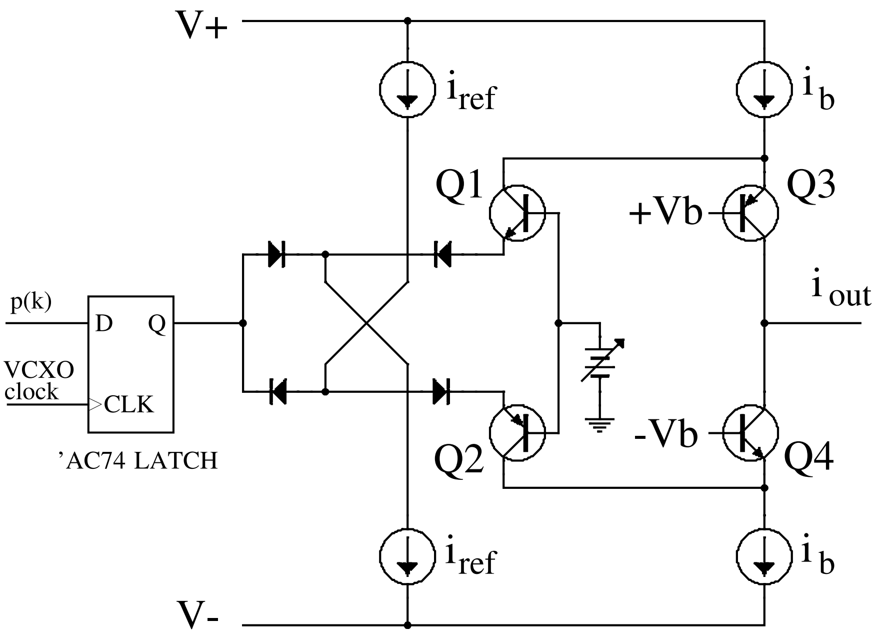

Figure C.8: The implemented NRZ current switch one-bit DAC. The output current is either $+i_{\mathrm{ref}}$ or $-i_{\mathrm{ref}}$ depending on the digital input code $p(k)$. The adjustable voltage source adjusts the skewness of the low to high and high to low transitions relative to the clock.

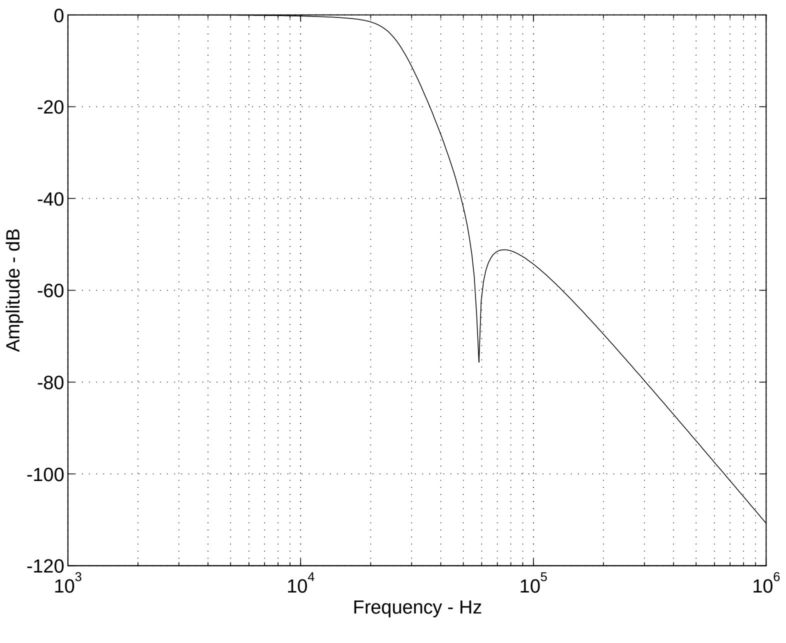

Figure C.9: Amplitude response for the analog post filters.

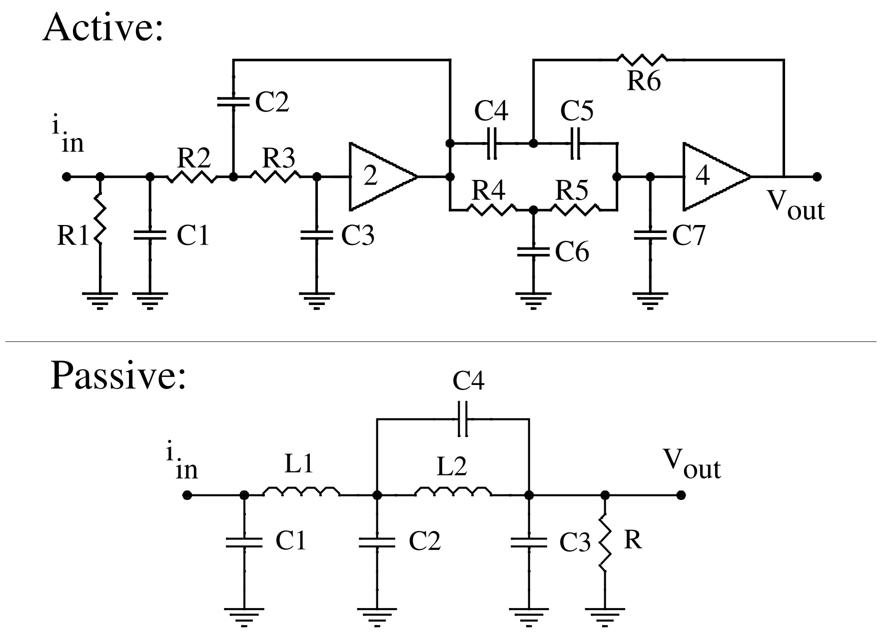

Figure C.10: The implemented passive and active analog post filters.

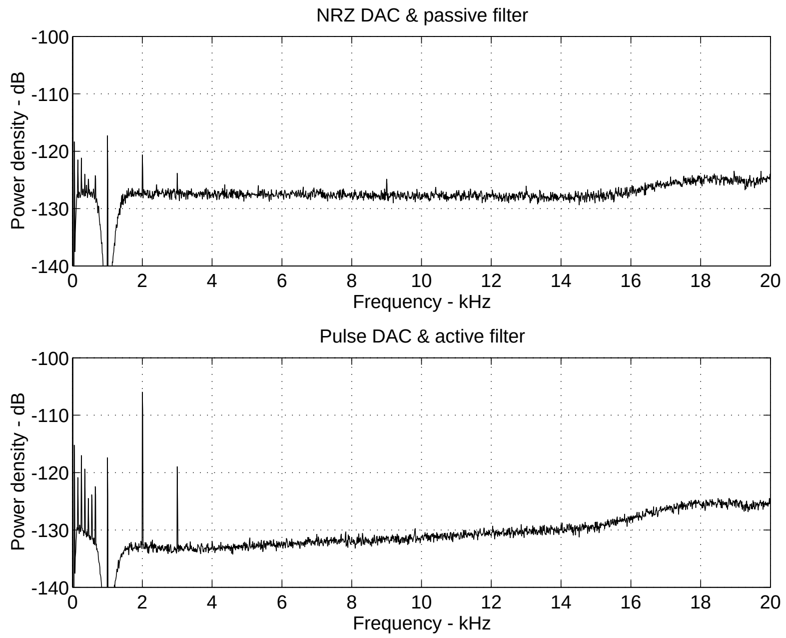

Figure C.11: Output spectra of the entire converter with 0dB 1kHz sinewave input quantized to 18 bit using triangular dither. The spectra are measured using a notch filter and 64 times averaged 16k FFTs at sample rate 44.1 kHz. Upper graph: NRZ DAC and passive post filter. Lower graph: Pulse DAC and active filter.

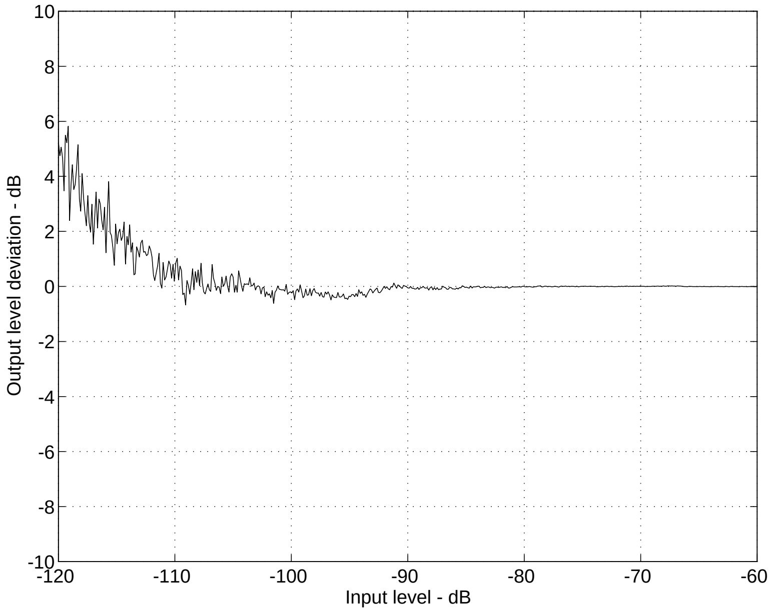

Figure C.12: Level linearity of the pulse type DAC and active filter with 18-bit triangular dithered sine input at 500 Hz.

Bibliography

[1] R. W. Adams *et. al.,* "Theory and practical implementation of a 5th-order Sigma-Delta A/D converter," *J. Audio Eng. Soc.,* vol. 39, pp. 515–528, July/Aug. 1991.

[2] R. W. Adams, *private communication*, Nov. 1992.

[3] B. P. Agrawal, K. Shenoi, "Design methodology for ΣΔM," *IEEE Trans. on Communication*, vol. 31, no. 3, pp. 360–369, March 1983.

[4] D. Anastassiou, "Error diffusion coding for A/D conversion," *IEEE Trans. on Circuits and Systems*, Vol. 36, No. 9, pp. 1175–1186, Sept. 1989.

[5] S. H. Ardalan and J. J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Sys.*, vol. CAS–34, pp. 593–603, June 1987.

[6] T. Berger, *Rate distortion theory — a mathematical basis for data compression*, Prentice-Hall, 1971.

[7] G. Burra, K. S. Chao, "A high-speed high-resolution oversampled A/D converter," *IEEE Proc. ISCAS 1993*, pp. 1282–1285, May 1993.

[8] J. C. Candy, "A use of double integration in sigma-delta modulation," IEEE Trans. on Communications, vol. 33, no. 3, pp. 249–258, March 1985.

[9] J. C. Candy, G. C. Temes, eds., *Oversampled delta-sigma data converters. Theory, design and simulation*, IEEE Press, 1992.

[10] K. C.-H Chao, S. Nadeem, W. L. Lee and C. G. Sodini, "A higher order topology for interpolative modulators for oversampling A/D conversion," *IEEE Trans. Circuits. Sys.*, vol. CAS–37, pp. 309–318, Mar. 1990.

[11] L. O. Chua, T. Lin, "Chaos in digital filters," *IEEE Trans. Circuits and Systems*, vol. 35, no. 6, pp. 648–658, June 1988.

[12] W. Chou, R. M. Gray, "Dithering and its effect on sigma-delta and multistage sigma-delta modulation," *IEEE Trans. on Information Theory*, vol. 37, no. 3, pp. 500–513, May 1991.

[13] P. Cvitanović, "Chaos for cyclists," printed in: *Noise and chaos in nonlinear dynamical systems,* F. Moss, L. Lugiato and W. Schleich, eds. Cambridge Univ. Press, 1990, pp. 270–288.

[14] C. Dunn and M. O. J. Hawksford, "Is the AESEBU/SPDIF Digital Audio Interface Flawed?," *Proc. 93rd AES Convention*, Preprint # 3360, Oct. 1992.

[15] O. Feely, L. O. Chua, "Nonlinear dynamics of a class of analog-to-digital converters," *Int. journal of Bifurcation and Chaos*, Vol. 2, No. 2, pp. 325–340, 1992

[16] O. Feely, L. O. Chua, "The effect of integrator leak in Σ-Δ modulation," *IEEE Trans. on Circuits and Systems*, Vol. 38, No. 11, pp. 1293–1305, Nov. 1991.

[17] I. Fujimori, K. Hamashita, E. J. Swanson, "A fifth-order delta-sigma modulator with 110 dB audio-band dynamic range," *Proc. 93rd AES Convention*, Preprint # 3415, Oct. 1992.

[18] R. A. Gabel, R. A. Roberts, *Signals and linear systems*, Wiley, 1987.

[19] A. Gersho, R. M. Gray, *Vector quantization and signal compression*, Kluwer Academic Publishers, 1992.

[20] M. Gerzon, P. G. Craven, "Optimal noise shaping and dither of signals," *Proc. 87th AES Convention*, Preprint # 2822, Oct. 1989.

[21] A. Grace, *Optimization TOOLBOX – user's guide*, The Mathworks Inc, 1992.

[22] C. Grebogi, E. Ott and J. A. Yorke, "Chaotic attractors in crisis," *Phys. Rev. Lett.*, vol. 48, pp. 1507–1510, 1982.

[23] S. R. Green, S. Harris and B. Wilson, "An 18-bit Delta-Sigma D/A Processor System Achieving Full-Scale THD+N>100dB," *Proc. 93rd AES Convention*, Preprint # 3416, Oct. 1992.

[24] S. Harris, "How to achieve optimum performance from Delta-Sigma A/D and D/A converters," *J. Audio Eng. Soc.*, vol. 41, no. 10, pp. 782–789, Oct. 1992.

[25] S. Hein and A. Zakhor, *Sigma-Delta modulators – nonlinear decoding algorithms and stability analysis*, Kluwer Academic Publishers, 1993.

[26] S. Hein, "Exploiting chaos to suppress spurious tones in general double-loop Σ-Δ modulators," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 10, pp. 651–659, Oct. 1993.

[27] J. Isoaho, H. Tenhunen, J. Heikkilä and L. Lipasti, "High Resolution DAC Design based on FPGAs," *Proc. Oxford 1991 Int. Workshop on FPGA Logic and Applications*, pp. 343–352, Sep. 1991.

[28] T. Karema, T. Ritoniemi and H. Tenhunen, "Intermodulation in Sigma-Delta D/A Converters," *Proc. IEEE ISCAS'91*, pp. 1625–1628.

[29] J. G. Kenney, L. R. Carley, "Design of multibit noise-shaping data converters," *Analog Integrated Circuits and Signal Processing*, 3, pp. 99–112, 1993.

[30] C. Knudsen, *Aspects of noninvertible dynamics and chaos,* Ph.D. dissertation, Physics Department, The Technical University of Denmark, Jan. 1994.

[31] B. M. J. Kup, E. C. Dijkmans, P. J. A. Naus, J. Sneep, "A bitstream digital-to-analog converter with 18-b resolution," *IEEE Journal of Solid-state Circuits*, vol. 26, No. 12, pp. 1757–1763, Dec. 1991.

[32] J. L. LaMay, H. T. Bogard, "How to obtain maximum practical performance from state-of-the art delta-sigma analog-to-digital converters," *IEEE Trans. on Instrumentation and Measurements*, vol. 41, no. 6, pp. 861–867, December 1992.

[33] R. C. Ledzius, J. Irwin, "The basis and architecture for the reduction of tones in a sigma-delta DAC," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 7, pp. 429–439, July 1993.

[34] W. L. Lee & C. Sodini, "A topology for higher order interpolative coders," *IEEE Proc. ISCAS'87*, pp. 459–462.

[35] J. N. Little and L. Shure, *Signal processing TOOLBOX - user's guide*, The MathWorks, Inc., 1993.

[36] E. Meitner and R. Gendron, "Time Distortions Within Digital Audio Equipment Due to Integrated Circuit Logic Induced Modulation Products," *Proc. 91st AES Convention*, Preprint # 3105, Oct. 1991.

[37] M. Motamed, A. Zakhor and S. Sanders, "Tones, saturation and SNR in double loop $\Sigma - \Delta$ modulators," *Proc. IEEE ISCAS'93,* pp. 1345–1348, Chicago, USA, May 1993.

[38] P. J. A. Naus *et. al.*, "A CMOS Stereo 16-bit D/A Converter for Digital Audio," *IEEE J. Solid-State Circuits*, vol. SC–22, pp. 390–394, June 1987.

[39] Truong-Thao Nguyen, *Deterministic analysis of oversampled A/D conversion and sigma-delta modulation, and decoding improvements using consisten estimates*, Ph.D. dissertation, Center for Telecommunications Research, Columbia University, New York, 1993.

[40] S. R. Norsworthy and D. A. Rich, "Idle channel tones and dithering in Delta-Sigma modulators," *Proc. 95th AES Convention*, preprint # 3711, Oct. 1993.

[41] S. R. Norsworthy, "Optimal nonrecursive noise shaping filters for oversampling data converters," *IEEE Proc. ISCAS'93,* pp. 1353–1356, Chicago, USA, May 1993.

[42] S. R. Norsworthy, "Effective dithering of sigma-delta modulators," *IEEE Proc. ISCAS 1992*, pp. 1304–1307, May 1992.

[43] T. Okamoto, Y. Maruyama, A. Yukawa, "A stable high-order delta-sigma modulator with an FIR spectrum distributor," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 7, pp. 730–735, July 1993.

[44] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*, Prentice Hall, 1989.

[45] S. J. Park, R. M. Gray, "Sigma-Delta modulation with leaky integration and constant input," *IEEE Trans. on Information Theory*, Vol. 38, No. 4, pp. 1512–1533, Sep. 1991.

[46] T. S. Parker, L. O. Chua, *Practical numerical algorithms for chaotic systems*, Springer Verlag, 1989.

[47] S. C. Pinault, P. V. Lopresti, "On the behavior of the double-loop sigma-delta modulator," *IEEE Trans. on Circuits and Systems –II*, Vol. 40, No. 8, pp. 467–479, Aug. 1993.

[48] L. Risbo, "FPGA based 32 times oversampling 8th order Sigma-Delta audio DAC," *Proc. 96th AES Convention*, preprint # 3808, Feb. 1994, Reprinted in Appendix C.

[49] L. Risbo, *Synthesis and simulations of stable high-order Sigma-Delta modulators*, M.sc. Thesis (in Danish), Electronics Institute, Technical University of Denmark, Aug. 1991.

[50] L. Risbo, *Stability analysis of a class of chaotic second-order* Σ-Δ *modulators*, Report prepared in course 1751 (in Danish), MIDIT, Technical University of Denmark, June 1993.

[51] L. Risbo, "Stability predictions for high-order Sigma-Delta modulators based on quasilinear modeling," *IEEE Proc. ISCAS'94,* pp. 5.361–5.364, London, UK, May 1994.

[52] L. Risbo, "Improved stability and performance from Σ-Δ modulators using one-bit vector quantization," *IEEE Proc. ISCAS'93,* pp. 1365–1368, Chicago, USA, May 1993, Reprinted in Appendix B.

[53] L. Risbo, "On the design of tone free Σ-Δ modulators," Preliminary manuscript, march 1994, Accepted for *IEEE Trans. on Circuits and Systems –II*, Reprinted in Appendix A.

[54] L. Risbo, *This Dissertation.*

[55] T. Ritoniemi, T. Karema and H. Tenhunen, "Design of Stable High Order Sigma-Delta Modulators," *Proc. IEEE ISCAS'90*, pp. 3267–3270, May 1990.

[56] T. Ritoniemi, T. Karema, H. Tenhunen, "Modelling and performance estimation of sigma-delta modulators," *IEEE Proc. ISCAS 1991*, pp. 2705–2708, May 1991.

[57] R. Schreier, "Destabilizing limit cycles in Delta-Sigma modulators with chaos," *IEEE Proc. ISCAS'93,* pp. 1369–1372, Chicago, USA, May 1993.

[58] R. Schreier, "An empirical study of high-order single-bit delta-sigma modulators," *IEEE Trans. Circuits and Syst.—II: Analog and Digital Signal Processing*, vol. 40, no. 8, pp. 461–466, Aug. 1993.

[59] R. Schreier, Y. Yang, "Stability tests for single-bit sigma-delta modulators with second-order FIR noise transfer functions," *IEEE Proc. ISCAS 1992*, pp. 1316–1319, May 1992.

[60] R. Schreier, M. Snelgrove, "Stability in a general Σ-Δ modulator," *IEEE Proc. ICASP 1991*, vol. 3 pp. 1769–1772, May 1991.

[61] N. S. Sooch, J. W. Scott, T. Tanaka, T. Sugimoto and C. Kubomura, "18-bit Stereo D/A Converter with Integrated Digital and Analog Filters," *Proc. 91st AES Convention*, Preprint # 3113, Oct. 1991.

[62] H. A. Spang, P. M. Schultheiss, "Reduction of quantization noise by use of feedback," *IRE Trans. on Communications Systems*,pp. 285–317, Dec. 1962.

[63] E. F. Stikvoort, “Some remarks on the stability and performance of the noise shaper or sigma-delta modulator,” IEEE Trans. on Communications, vol. 36, no. 10, pp. 1157–1162, Oct. 1988.

[64] J. T. Stonick, S. H. Ardalan, J. K. Townsend, “An improved analysis of sigma-delta modulators with bandlimited Gaussian input,” *IEEE Proc. ICASSP’90*, pp. 1691–1694, 1990.

[65] J. T. Stonick, J. L. Rulla, S. H. Ardalan, J. K. Townsend, “A new architecture for second order $\Sigma\Delta$ Modulation,” *IEEE Proc. ISCAS’90*, pp. 360–363, 1990.

[66] S. K. Tewksbury, R. W. Hallock, “Oversampled, linear predictive and noise-shaping coders for order $N > 1$,” *IEEE Trans. Circuits and Systems*, vol. 25, no. 7, pp. 436–447, July 1978.

[67] H. Wang, “A geometric view of $\Sigma\Delta$ modulation,” *IEEE Trans. on Circuits and Systems –II*, Vol. 39, No. 6, pp. 402–405, June 1992.

[68] H. Wang, *A study of sigma-delta modulations as dynamical systems*, Ph.D. dissertation, Center for Telecommunications Research, Columbia University, New York, 1993.

[69] C. Wolff, L. R. Carley, “Modeling the quantizer in higher-order delta-sigma modulators,” *IEEE Proc. ISCAS’88,* pp. 2335–2338, Helsinki, Finland, June 1988.

Dansk Resumé

Σ-Δ modulation anvendes i udstrakt grad i A/D og D/A omsættere til krævende audio formål. Denne modulationsform indkoder et tidsdiskret signal med høj opløsing om til et to-niveau (dvs. et-bit) signal med en samplingfrekvens væsentligt højere end krævet af sampling sætningen (oversampling). Herved opnås en høj amplitudeopløsning på bekostning af lavere opløsning i tid. Denne virkning er bedre kendt som "noise-shaping", dvs. fejlen ved et-bit signalet er fjernet fra et lavfrekvent bånd og koncentreret ved høje frekvenser.

En Σ-Δ modulator består af en et-bit kvantisator indeholdt i en modkoblingssløjfe med et lineært filter. Modulatoren er dermed et ulineært dynamisk system. For nogle typer filtre bliver modulatoren tillige kaotisk. En alvorlig ulempe ved Σ-Δ modulatoren er muligheden for ustabilitet. Dette er specielt et praktisk problem for modulatorer med høj filterorden.

Stabilitetsproblemet er hovedemnet for denne afhandling. Problemet behandles i første omgang fra et system-dynamisk synspunkt. En af konlusionerne er, at stabiliteten af kaotiske modulatorer forsvinder ved en såkaldt *boundary crisis*, der opstår, når attraktoren kolliderer med sit attraktionsbasin. Graden af ustabilitet kan beskrives ved den såkaldte escape-rate.

Stabilitet kan også analyseres vha. symbolsk dynamik. Konlusionen er, at stabiliteten er tæt knyttet til antallet af mulige grænsecykler.

Nogle modulatorer bliver karakteriseret som værende upålidelige. Dette vil sige, at ustabiliteten ikke indtræffer ved et veldefineret punkt. Sådanne modulatorer kan være tilsyneladende stabile ved kortere simuleringer. Ustabiliteten viser sig kun ved enten meget lange simuleringer eller ved realtidsimplementeringer.

Afhandlingen præsenterer værktøjer til optimering af både signal/støj forhold og pålidelighed. Metoden kombinerer eksakt analyse med approksimativ lineariseret analyse. Metoden har været anvendt til design af en 32 gange oversamplende 8. ordens Σ-Δ D/A-omsætter til audio brug. Denne omsætter er implemeteret vha. Field Programmable Gate Array kredse.

Afhandlingen behandler også metoder til undertrykkelse af de meget kraftige toner omkring den halve sampling frekvens. Brugen af kaos og dither-støj viser sig at være nogenlunde lige effektive og ækvivalente på flere måder. Med hensyn til brug af kaotiske modulatorer, demonstreres en metode, hvor støjoverføringsfunktionen forlænges med et alpasled.

Til slut præsenteres en ny klasse af modulatorer, der anvender et-bit vektor kvantisering. Det vises, at introduktionen af en vektorkvantisator kan forbedre stabiliteten betydeligt. Samtidig kan styrken af tonerne ved den halve samplingfrekvens reduceres.