Improved stability and performance from sigma-delta modulators using 1-bit vector quantization

Risbo, Lars

Published in:
Proceedings of the IEEE International Symposium on Circuits and Systems

Link to article, DOI:
10.1109/ISCAS.1993.393985

Publication date:
1993

Document Version
Publisher's PDF, also known as Version of record

Improved Stability and Performance from Sigma-Delta Modulators using 1-bit Vector Quantization

Lars Risbo
Electronics Institute, Technical University of Denmark
Bld. 349, DK-2800 Lyngby, Denmark,
e-mail : lrisbo@eiffel.ei.dth.dk

Abstract - In this paper, a novel class of Sigma-Delta modulators is presented. The usual scalar 1-bit quantizer in a Sigma-Delta modulator is replaced by a 1-bit vector quantizer with a N-dimensional input state-vector from the linear feed-back filter. Generally, the vector quantizer changes the nonlinear dynamics of the modulator, and a proper choice of vector quantizer can improve both system stability and coding performance. The paper shows how to construct the vector quantizer in order to limit the excursions in state-space. The proposed method is demonstated graphically for a simple second order modulator.

I. INTRODUCTION

The use of Sigma-Delta modulators in the construction of high resolution A/D and D/A converters has in recent years become widespread. Sigma-Delta modulators perform a coding of a bandlimited signal using an extremely simple alphabet composed of only two levels (1-bit). The theoretical task of 1-bit signal coding can be described as a search for a path close to the input signal in an (infinite) binary tree representing the set of 1-bit signals. The resulting 1-bit digital signal approximates the input signal within a narrow base-band, and it contains a large amount of quantization noise power concentrated outside the base-band. This phenomenon is commonly known as "noise-shaping". In order to obtain good coding performance the 1-bit signal coding must be done using a high degree of oversampling, i.e., the sampling frequency must be considerably higher than usually prescribed by the sampling theorem.

A Sigma-Delta modulator is a feed-back loop composed of a 1-bit quantizer (sign detector or signum function) and a linear feed-back filter. The choice of feed-back filter is crucial for the modulator performance. The use of high order filters enables more efficient coding with a reduced demand for oversampling, and the quantization noise becomes less dependent on the input signal. Unfortunately, high-order Sigma-Delta modulators suffer from stability problems which constitute a severe disadvantage in practical implementations.

The purpose of this paper is to introduce a new class of Sigma-Delta modulators employing a 1-bit vector quantizer in order to improve both coding performance and stability.

0-7803-1254-6/93$03.00 © 1993 IEEE

II. TRADITIONAL SIGMA-DELTA MODULATION

A block diagram of a Sigma-Delta modulator with conventional scalar quantization is shown in Figure 1 (if the vector quantizer marked VQ is disregarded). The output signal, y_k, is subtracted from the input signal, i_k, and fed into the feed-back filter, which usually is a low-pass filter. The resulting output signal, e_k, is an error signal indicating the instantaneous coding error with emphasis on the base-band. The output signal is a 1-bit quantization of this error signal and, in normal stable operation the modulator tracks the input signal, thus keeping the amplitude of the error signal low.

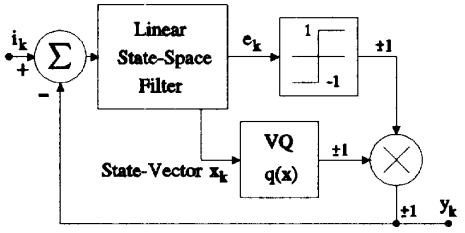

Figure 1 Block diagram of a Sigma-Delta modulator with a vector quantizer (VQ).

The dynamics of a N'th order modulator can be described in a compact form by the following state-space model:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{b}(i_k - y_k) \\ e_k &= [1, 0, \ldots, 0]\boldsymbol{x}_k \\ y_k &= \mathrm{sgn}(e_k) \end{aligned} \qquad (1)$$

where $\boldsymbol{x}_k$ is a N-dimensional state vector corresponding to time step k, $\boldsymbol{A}$ is the $N \times N$ transition matrix for the feed-back filter determining the filter poles, and $\boldsymbol{b}$ is a N-dimensional feed-back vector determining the filter zeros. Usually the poles of the feed-back filter are specified at first for maximum noise suppression in the base-band, and afterwards the zeros of the filter are adjusted for best trade off between stability and SNR. The error signal, e_k, is for simplicity defined as the first state vector coordinate (without loss of

generality). The signum function sgn(·) maps positive arguments into 1 and gives -1 otherwise.

The traditional Sigma-Delta modulator represents a tree search algorithm, in which the 1-bit quantizer continuously picks the branch yielding the least magnitude of the error signal. This local optimization leads normally to a well-behaved error signal, e_k. However, local optimization may under various circumstances fail to track the input signal and thus cause the modulator to become unstable. The instability is characterized by an oscillating error signal with large amplitude and low frequency. The modulator locks into a limit cycle with finite amplitude provided that the poles of the feed-back filter (the eigenvalues of A) are inside the unit circle [1]. Poles outside the unit circle may result in an error signal diverging to infinity. The modulator has to be reset (i.e. setting $x_k=0$) in order to leave the unstable mode and resume tracking of the input signal. Generally, the set of initial state-vectors, x_0, can be divided into two disjoint sets giving stable or unstable behavior, respectively. The boundaries between these two sets are typically very complicated and may even be fractal.

Some practical implementations incorporate limiters in the feed-back filter in order to control the instability [2], and in other implementations the modulator is reset if the output starts oscillating [3]. Both solutions can lead to audible "clicks" in connection with audio systems. A more feasible solution is to design the feedback filter in such a way that the modulator stays stable if the input signal is kept within a certain amplitude range. Unfortunately there is currently no analytical tool which can tell precisely when instability arises, and the design process normally relies on extensive simulations. In general, the modulator design is a trade off between stability and coding performance.

III. SIGMA-DELTA MODULATION EMPLOYING VECTOR QUANTIZATION

The proposed new class of Sigma-Delta modulators (Figure 1) contains a supervising 1-bit vector quantizer which inverts the modulator output if the filter state vector enters certain critical regions. This represents a tree search algorithm which occasionally departs from blind local optimization leading to possibly better global properties. It should be emphasized that the operation of the linear feedback filter is unaffected, i.e. the error signal, e_k, still represents the base-band coding error. Generally, it is a very complicated task to construct a suitable vector quantizer. Consequently, the following section will only describe a certain class of vector quantizers which aim at limiting the error signal. A limitation of the error signal will obviously insure stability and may possibly also reduce the base-band coding error.

A set of "legal" state vectors giving a limited error signal e_k is defined by means of two bounds, e_{min} and e_{max}:

$$\begin{aligned} L&=\{x\in\mathbb{R}^N \mid e_{\min}<e<e_{\max}\} \\ &=\{x\in\mathbb{R}^N \mid e_{\min}<[1,0,...,0]x<e_{\max}\} \end{aligned} \tag{2}$$

The intention is, that the modulator state-vector never should leave this arbitrarily chosen area in state-space.

The two maps $\mathcal{F}_+$, $\mathcal{F}_-$: $\mathbb{R}^N\mapsto\mathbb{R}^N$ map a state vector from time step k to $k+1$. $\mathcal{F}_+$ uses normal scalar 1-bit quantization and $\mathcal{F}_-$ uses the inverted quantizer output:

$$\begin{aligned} \mathcal{F}_+(x)&=Ax+b(i_k-\mathrm{sgn}([1,0,...,0]x)) \\ \mathcal{F}_-(x)&=Ax+b(i_k+\mathrm{sgn}([1,0,...,0]x)) \end{aligned} \tag{3}$$

The nonlinear dynamics of the modulator can be studied by iterating these maps on an initial state vector. Note that the input signal, i_k, in (3) is treated like a known parameter.

Consider the following class of sets of state vectors, G_n, which stay in L during the next n (normal) time steps using $\mathcal{F}_+$:

$$\begin{aligned} G_0&\triangleq L \\ G_{n+1}&\triangleq \mathcal{F}_+^{-1}(G_n)\bigcap L \end{aligned} \tag{4}$$

Another class of sets, H_n, contains legal state vectors which are mapped outside L after precisely n usual $\mathcal{F}_+$ time steps:

$$\begin{aligned} H_0&\triangleq \mathbb{R}^N\setminus L \\ H_{n+1}&\triangleq \mathcal{F}_+^{-1}(H_n)\bigcap L \end{aligned} \tag{5}$$

It can be concluded from the above mentioned properties of the H_n sets, that these sets are disjoint.

The purpose of the 1-bit vector quantizer is to identify regions in state space where modulator output inversion ensures that the error signal stays within the legal range during the succeeding time steps. Consider the following class of sets:

$$F_n\triangleq \mathcal{F}_+^{-1}(H_{n-1})\bigcap \mathcal{F}_-^{-1}(G_{n-1}) \tag{6}$$

The elements in F_n represent state vectors which are mapped outside L after precisely n usual $\mathcal{F}_+$ time steps, and in addition the vectors stay in L after a $\mathcal{F}_-$ time step followed by n-1 normal $\mathcal{F}_+$ time steps. Since the H_n -sets are disjoint, it is easy to conclude that the F_n sets are disjoint too.

The 1-bit vector quantizer in the proposed modulator architecture will invert the output of the scalar quantizer if the feed-

back filter state vector belongs to one of the first N_f F_n -sets. The state-space model (1) can thus be modified to:

$$\begin{aligned} x_{k+1} &= Ax_k + b(i_k - y_k) \\ e_k &= [1, 0, ..., 0]x_k \\ y_k &= q(x_k)\,\mathrm{sgn}(e_k) \end{aligned} \qquad (7)$$

$$q(x) \triangleq \begin{cases} 1 & , x \notin \bigcup_{n=1}^{N_f} F_n \\ -1 & , x \in \bigcup_{n=1}^{N_f} F_n \end{cases}$$

The N_f parameter indicates the time horizon of the prediction built into the supervising vector quantizer. Generally, the system in (7) is not guaranteed to yield an error signal within the legal range. If the e_{min} and e_{max} bounds are too tight in combination with a specific feed-back filter and input signal, no 1-bit signal having an error signal within the legal range exists. Under these circumstances the system may either become totally unstable or stable with occasional excursions away from L. Consequently, the error bounds, e_{min} and e_{max}, and N_f should be chosen carefully in order to insure proper operation. In addition, notice that the vector quantizer could also be derived from other and more complex definitions of the set of legal state-vectors, L.

The pseudo C-code function F-test shown i Figure 2 can be used in numerical simulations for identification of the F_n sets. The piecewise linearity of the $\mathcal{T}_+$ and $\mathcal{T}_-$ maps combined with the choice of legal state vectors, L, results in F_n sets delimited by hyperplanes in state space with at most n different normal vectors. A proper choice of rotated coordinates for the state-space transforms the F_n-sets into rectangular boxes. This enables the vector quantizer to be constructed from an ensemble of scalar comparators operating on N_f auxiliary outputs from the state-space filter followed by some Boolean operations. Consequently, the modulator with vector quantization could be interpreted as a modulator with N_f different feed-back filters and comparators.

```
function F-test(x)

// Returns n if (x∈Fn ∧ n≤Nf)        //
// and 0 otherwise                   //
{
  x' = x; n = 1;
  x = T+(x);
  x' = T+(x');
  while (x∈L ∧ x'∈L ∧ n<Nf)
  {
        x = T+(x);
        x' = T+(x');
        n++;
  }
  if (x∉L ∧ x'∈L) Return(n)
  else Return(0);
}
```

Figure 2 Pseudo C-code for the function F-test.

IV. SIMULATION EXAMPLE

The proposed method can be demonstrated graphically for a second order modulator. The following parameters represent a feed-back filter composed from two cascaded first order discrete-time integrators:

$$A = \begin{bmatrix} 1 & 1 \\ -a & 1 \end{bmatrix}, \; a = 0.005, \qquad b = [b_1, b_2]^T = [1, 1]^T \qquad (8)$$

The a-coefficient introduces local feed-back around the integrators and locate the filter poles, $\lambda = 1 \pm j\sqrt{a}$, outside the unit circle. The traditional modulator according to (1) is intentionally unstable with $b = [1,1]^T$, resulting in an infinite error signal even with zero input (the modulator becomes stable if b_2 is reduced to 0.5). Figure 3 shows a simulation of this (normally) unstable modulator in state-space using a vector quantizer with parameters $N_f = 2$, $e_{max} = 2.6$, $e_{min} = -1.2$ and DC input arbitrarily set to $i_k = 0.44$. The F_n-sets are shown in different shadings and 10^5 simulated state vectors are plotted as inverted color. It is seen that the introduction of this fairly primitive vector quantizer is sufficient for gaining stability, even though the error signal exceeds the e_{max} boundary slightly ($e_k < 3.48$). The vector quantizer inverts the modulator output in 5.4 % of the time steps.

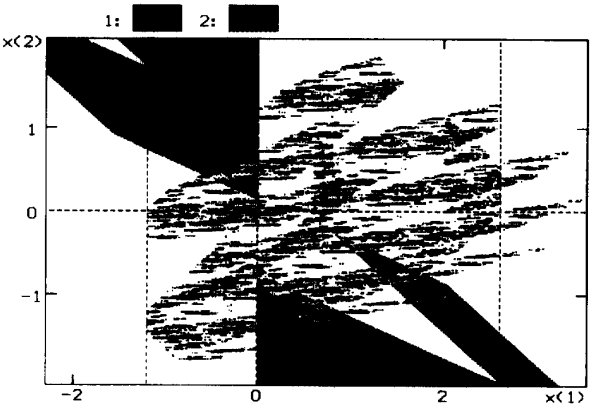

Figure 3 Simulation with $N_f = 2$, $i_k = 0.44$, $e_{max} = 2.6$, $e_{min} = -1.2$, 10^5 time steps.

In Figure 4 N_f is raised to 4 leading to a more complex vector quantizer. Now the error signal is totally bounded by e_{max} and e_{min}. Notice also the visual change in the dynamics of the system. The scalar quantizer output is inverted in 7.8 % of the time steps.

Besides the gain in stability the introduction of the vector quantizer may also improve coding performance. This is due to the limitation of the error signal. Figure 5 shows power spectra of the modulator output from two simulations with

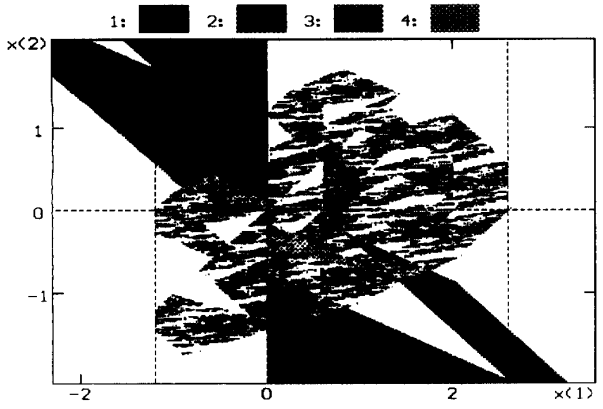

Figure 4 Simulation with $N_f=4, i_k=0.44, e_{max}=2.6, e_{min}=-1.2$, 10^5 time steps.

different pairs of error bounds. It is seen that a tightening by a factor 10 of the error bounds leads to 25 dB reduction in low frequency coding error. This enormous improvement is not a general figure but rather an indication of a poor choice of the feed-back vector ***b***. Generally, the vector quantizer and the feed-back vector should be optimized simultaneously for the best result.

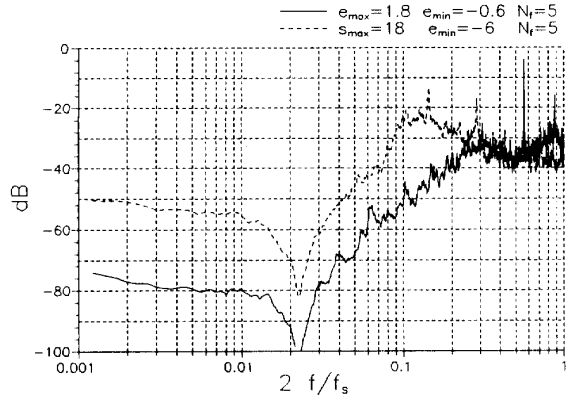

Figure 5 Power spectra of y_k for two simulations (8) with DC input $i_k=0.44$ (10^6 time steps, 8k FFT, Kaiser-Bessel window).

V. VECTOR QUANTIZATION APPLIED TO HIGH-ORDER MODULATORS

Simulations with high-order ($N>2$) modulators employing vector quantization show that stability can be improved considerably. Typically the input amplitude range for stable operation can be extended significantly and modulators which normally are unstable can be stabilized. The evident improvement in coding performance seen for the second order modulator with tight error bounds is not so clear for high-order modulators. Too tight error bounds lead to a decrease in the coding performance and possible instability. The gain in coding performance using vector quantization is also highly dependent on the choice of feed-back filter. The extension of the reliable and stable amplitude range is, however, sufficient for an increased overall SNR performance. A similar gain in performance can typically be obtained with a carefully optimized feed-back filter with increased order without vector quantization. Therefore, the added circuit complexity from a vector quantizer should be compared to an increased filter order in any case.

An eighth-order modulator intended for digital audio with 32 times oversampling was used for evaluation of the benefits from vector quantization. The unmodified modulator has a maximum input amplitude of approx. 0.2 (relative to full scale) for reliable and stable operation. It was possible to operate the modulator with both DC and sinusoidal inputs with amplitudes up to approx. 0.5 using a vector quantizer with $N_f=35$. Only a slight increase in the base-band noise power was observed outside the normally stable amplitude range.

VI. CONCLUSION

A novel class of Sigma-Delta modulators employing vector quantization has been presented. The vector quantizer acts as a reflecting barrier which limits the excursions in state-space without introducing transient increase in the coding error. This can significantly extend the usable amplitude input range for high-order modulators leading to better SNR figures. The price paid is an increased circuit complexity stemming from the - generally - signal dependent vector quantizer. For the purpose of real-time implementations various simplifications of the vector quantizer should be investigated.

REFERENCES

[1] Søren Hein & Avideh Zakhor, "On the stability of interpolative Sigma Delta modulators," Proc. IEEE ISCAS-91, pp. 1621-1624, June 1991.

[2] E. F. Stikvoort, "Some remarks on the stability and performance of the Noise Shaper or Sigma Delta modulator," IEEE Trans. Comm., vol. 36, no. 10, pp. 1157-1162, Oct. 1988.

[3] R.W. Adams, P.F. Ferguson Jr., S. Vincellette, A. Ganesan, T. Volpe, B. Libert, "Theory and practical implementation of a 5th-order Sigma-Delta A/D converter," 90th Audio Eng. Soc. Convention, Feb. 1991, Preprint # 3017.