RATE-DISTORTION ANALYSIS OF DELTA-SIGMA MODULATORS

Shuichi Ohno[⋆]*, Teruyuki Shiraki*[⋆]*, M.Rizwan Tariq*[⋆]*, and Masaaki Nagahara*[†]

[⋆]Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, 739-8527, JAPAN
[†] The University of Kitakyushu, Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, JAPAN

ABSTRACT

Delta-Sigma modulators are often utilized to convert analog signals into digital signals. The quantization error of a Delta-Sigma modulator can be reduced by oversampling. However, oversampling increases the number of bits per time if the same number of bits are assigned to each output of the quantizer. Due to limited bandwidth, the rate-distortion relation is necessary to balance the rate and the distortion. In this paper, we analyze the relationship between the rate and the distortion of an optimal scalar Delta-Sigma modulator that minimizes the variance of the error in the output of the system connected to the Delta-Sigma modulator. Numerical examples are provided to show rate-distortion relations of the optimal Delta-Sigma modulators.

***Index Terms*—** Quantization, Delta-Sigma modulator, rate-distortion

1. INTRODUCTION

Quantization is a fundamental process in signal processing. The simplest type of quantizer is the uniform quantizer, which has fixed-length codewords. The uniform quantizer is not efficient since it does not take account into the statistics of the input and/or the information on the system connected to the quantizer. Additional information can be exploited to obtain good quantizers. Under the assumption that the quantization error is a white uniformly distributed random sequence, among quantizers having a fixed-length codewords, the Lloyd-Max quantizer is optimal in the sense that it minimizes the distortion due to the quantization error [1, Chap.9]. However, the probability density function of the input to the quantizer is required to construct the Lloyd-Max quantizer.

Quantization with error feedback is more efficient than the uniform quantizer. It has a uniform quantizer and an error feedback filter, in which the filtered error of the uniform quantizer is fed back. Quantization with error feedback is adopted in Delta-Sigma modulators, which are often utilized to convert real values into fixed-point numbers and vice versa [2]. Error feedback filters have been designed to mitigate the quantization error [3, 4, 5, 6, 7]. Quantization with error feedback can also be used to reduce the effect of the quantized coefficients in fixed-point digital filters [8, 9].

It is known that when a Delta-Sigma modulator is used to quantize an analog signal into a digital signal, oversampling can reduce the error due to the quantization. However, oversampling increases the number of bits per time if the same number of bits are assigned to each output of the quantizer. It may degrade the distortion due to quantization if the number of bits per time is fixed. To balance the rate and the distortion, the rate-distortion relation of the Delta-Sigma modulator is necessary.

It has been found in [10] that for bandlimited signals, the variance of the distortion of a simple single-loop one-bit Delta-Sigma modulator decays at a rate of $O(\lambda^{-4})$, where λ is the oversampling ratio. In [11], it is proven that for bandlimited bounded signals, the squared maximum absolute value of the distortion of a one-bit Delta-Sigma modulator can decrease at a rate of $O(\lambda^{-4})$ and then a family of one-bit Delta-Sigma modulators that attain this rate has been provided. In [12], optimal filters in this family are designed to minimize the decay rate, which shows that an exponential rate of $O(2^{-0.102\lambda})$ is achieved by the designed filter. On the other hand, the mean squared error (MSE) of the optimal one-bit Delta-Sigma modulator that minimizes the MSE under the constraint on the variance of the input to the uniform quantizer decreases at an exponential rate of $O(2^{-0.807\lambda})$ [13]. This improvement becomes possible by exploiting the knowledge on the power spectral density function of the input, which is not always available, and by using additional pre-filter and post-filter with infinite orders. In this paper, we clarify the rate-distortion relation of conventional Delta-Sigma modulators without pre/post-filters when the spectrum of their input cannot be used.

After formulating our problem as an optimization problem, we show that the amplitude response of the optimal error feedback filter that minimizes the MSE can be parameterized by one parameter. The optimal error feedback filter can be determined numerically by minimizing the MSE with respect to this parameter. Then, the relationship between the number of bits used for quantization and the achievable MSE are clarified. This is our main contribution on the rate-distortion analysis of optimal Delta-Sigma modulators. It also demonstrates the contribution of oversampling to the reduction of the MSE. Numerical examples are provided to show the rate-distortion relation.

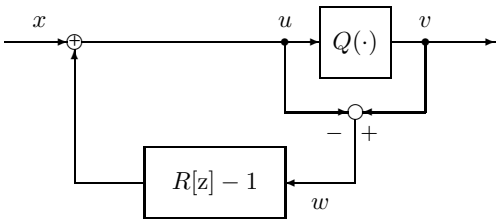

Fig. 1. Quantizer with an error feedback filter

2. QUANTIZATION WITH ERROR FEEDBACK

Figure 1 depicts our quantizer with error feedback, where x is the input to the quantizer with error feedback, v is its output, and $Q(\cdot)$ denotes a conventional static uniform quantizer. All signals are assumed to be of discrete-time. We denote the z transform of a sequence $f = \{f_k\}_{k=0}^{\infty}$ as $F[\mathrm{z}] = \sum_{k=0}^{\infty} f_k \mathrm{z}^{-k}$. We also express the output signal b of a linear time invariant (LTI) system, whose transfer function is $F[\mathrm{z}]$, to the input $a = \{a_k\}_{k=0}^{\infty}$ as $b = F[\mathrm{z}]a$, where z^{-1} is a unit-time delay operator.

In Fig. 1, the quantization error $w = v - u$ of the uniform quantizer is filtered by $R[\mathrm{z}] - 1$ and is fed back. The first coefficient of the impulse response of $R[\mathrm{z}]$ is assumed to be 1, which implies $R[\mathrm{z}] - 1$ is strictly causal. The minus 1 in $R[\mathrm{z}] - 1$ is just for simplicity of presentation.

The input u to the uniform quantizer is expressed as $u = x + (R[\mathrm{z}] - 1)w$. The quantization error of quantization with error feedback can be defined as $e = v - x$, which should be differentiated with the quantization error w of the uniform quantizer. It is easy to see that they are related such as $e = R[\mathrm{z}]w$. Then, the output of the quantizer can be expressed as

$$v = x + R[\mathrm{z}]w. \tag{1}$$

We assume that the output of the quantizer is the input to the system $P[\mathrm{z}]$ as depicted Fig. 2. The output y of $P[\mathrm{z}]$ can be expressed as $y = P[\mathrm{z}]v = P[\mathrm{z}]x + \epsilon$, where ϵ is the error in the output introduced by the quantization given by

$$\epsilon = P[\mathrm{z}]R[\mathrm{z}]w. \tag{2}$$

Quantization with error feedback has been developed to mitigate quantization errors in Delta-Sigma modulators [3, 4, 5, 6, 7] as well as in digital filters [8, 9]. If the frequency response of the input is available, then $R[\mathrm{z}]$, which is also called the *noise shaping filter*, can be designed to reduce the effect of w in the frequency band of x. This technique is known as *noise shaping* or *error spectrum shaping* [2, 3, 9, 14]. It has been shown in [13] that the mean squared error (MSE) of the optimal one-bit Delta-Sigma modulator decreases at an exponential rate of $O(2^{-0.807\lambda})$, where λ is the oversampling ratio. However, the input spectrum is often unavailable in practice. The purpose of this paper is to derive the MSE of Delta-Sigma modulators when the input spectrum cannot be used.

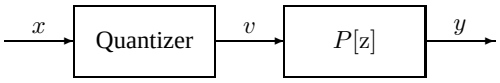

Fig. 2. Quantizer and system

3. RATE-DISTORTION ANALYSIS OF DELTA-SIGMA MODULATORS

The static uniform quantizer can be described by two parameters, the quantization interval $d(>0)$ and the saturation level $L(>0)$. Take a mid-rise quantizer as an example, whose output to an input ξ is expressed as

$$Q(\xi) = \begin{cases} \left(i + \frac{1}{2}\right) d, & \xi \in [id, (i+1)d), \ |\xi| \le L + \frac{d}{2} \\ L, & \xi > L + \frac{d}{2} \\ -L, & \xi < -L - \frac{d}{2} \end{cases} \tag{3}$$

for i being integer.

If we assign b bits to the mid-rise quantizer, then the number of quantization levels is 2^b. The dynamic range $[-L, L]$ of the mid-rise quantizer can be expressed as $2L = (2^b - 1)d$.

For our analysis, as in [13], we assume that a sufficient number of bits are assigned to the output of the uniform quantizer so that:

Assumption 1. *The error due to overloading (or equivalently, saturation) is negligible.*

The input x to the modulator is assumed to be a wide-sense stationary process having zero mean and variance σ_x^2. We also assume that:

Assumption 2. *The quantization error signal w of the uniform quantizer is a white random signal with zero-mean and variance σ_w^2 and uncorrelated with the input of the uniform quantizer.*

The dynamic range of the uniform quantizer is determined by the dynamic range of its input. It is reasonable to assume that [15] :

Assumption 3. *For a fixed number of quantization levels, the variance σ_w^2 of the quantization error of the uniform quantizer is proportional to the variance σ_u^2 of its input and the ratio is denoted as*

$$\gamma = \frac{\sigma_u^2}{\sigma_w^2}. \tag{4}$$

Let us denote the L_2 norm of a filter $H[\mathrm{z}]$ as $||H[\mathrm{z}]||$, which is defined as $||H[\mathrm{z}]|| = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} H^*[e^{j\omega}]H[e^{j\omega}]d\omega\right)^{\frac{1}{2}}$, where c^* is the complex conjugate of c.

From Assumption 2, the variance of the input to the uniform quantizer is expressed as

$$\sigma_u^2 = \sigma_x^2 + ||R[\mathrm{z}] - 1||^2 \sigma_w^2. \tag{5}$$

Then, under Assumption 3, the variance of the quantization error of the uniform quantizer is given by $\sigma_w^2 = \sigma_x^2/(\gamma - ||R[z]-1||^2)$, which requires $\gamma - ||R[z]-1||^2 > 0$. Since $R[z]$ has a unit gain, we have $||R[z]-1||^2+1 = ||R[z]||^2$ and

$$\sigma_w^2 = \frac{\sigma_x^2}{\gamma + 1 - ||R[z]||^2}. \tag{6}$$

The variance of the error in the output of the system introduced by the quantization is given by $||P[z]R[z]||^2\sigma_w^2$. Substituting (6) into this results in

$$||P[z]R[z]||^2\sigma_w^2 = \frac{||P[z]R[z]||^2}{\gamma + 1 - ||R[z]||^2}\sigma_x^2. \tag{7}$$

We have to obtain the minimum of the MSE given by (7) for a fixed number of bits. For given σ_x^2 and $P[z]$, we can minimize the MSE with respect to $R[z]$ as follows.

To stabilize the quantizer, $R[z]$ must be stable. Then, as σ_x^2 in (7) is a scalar, our problem can be formulated as the following minimization:

$$\min_{R[z]\in RH_\infty} \frac{||P[z]R[z]||^2}{\gamma + 1 - ||R[z]||^2} \tag{8}$$

subject to $R[\infty] = 1$ and

$$||R[z]||^2 < \gamma + 1 \tag{9}$$

where RH_∞ is the set of stable proper rational functions with real coefficients.

To enable theoretical analysis, we relax the stable proper rational function $R[z]$ to a function $r(\omega) \in L_2$, that is piecewise differentiable on $[-\pi, \pi]$, has at most a finite number of discontinuity points, and satisfies for $c_0 \geq 0$ that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln r(\omega) d\omega = c_0. \tag{10}$$

The L_2 norm of $q(\omega) \in L_2$ is defined as

$$||q(\omega)||^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} q^*(\omega)q(\omega)d\omega \tag{11}$$

We denote the set of L_2 functions that satisfies (10) as $\mathcal{C}_0$. We also define a subset of L_2 functions as

$$\mathcal{C}_1 = \left\{r(\omega) : ||r(\omega)||^2 < \gamma + 1\right\}. \tag{12}$$

Although we extend the class of functions, from Lemma 1 in [13], we can find a stable proper rational function $R[z]$ such that $|R[e^{j\omega}]|$ approximates $r(\omega)$ arbitrarily well on $[-\pi, \pi]$.

Now our problem is to find the optimal function such that

$$r_{opt}(\omega) = \arg \min_{r(\omega)\in\mathcal{C}_0\cap\mathcal{C}_1} \frac{||p(\omega)r(\omega)||^2}{\gamma + 1 - ||r(\omega)||^2}. \tag{13}$$

In the following, we omit the proofs for our results due to the lack of space, which are presented in [16].

The optimal function cannot be expressed in a closed-form but can be characterized with one parameter as follows:

Theorem 1. *For any $\gamma > 0$, the optimal function of* (13) *can be expressed with a parameter α as*

$$r_\alpha(\omega) = \frac{\theta(\alpha)}{\sqrt{p^2(\omega) + \alpha}} \tag{14}$$

where

$$\theta(\alpha) = \exp\left(\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln(p^2(\omega) + \alpha)d\omega\right). \tag{15}$$

Then, the optimal α can be found based on the following theorem:

Theorem 2. *For any $\gamma > 0$, the optimal α, denoted by α_{opt}, that minimizes the MSE, satisfies $\alpha_{opt} > 0$ and*

$$\gamma + 1 = \frac{\theta^2(\alpha_{opt})}{\alpha_{opt}}. \tag{16}$$

Now suppose that the discrete-time system $P[z]$ is the discretized version of the original continuous-time system $P(s)$, which is assumed to be bandlimited as follows:

Assumption 4. *The continuous-time system $P(s)$ is bandlimited in $[-\pi/T_s, \pi/T_s]$ and $1/T_s$ is its Nyquist frequency.*

Sampling with sampling period T_s/λ for λ a positive integer is known as oversampling. The integer λ is called the *oversampling ratio*, which is the sampling frequency divided by the Nyquist frequency. We assume that $P[z]$ is the sampled system of $P(s)$ with sampling period T_s/λ. To compare the MSE with the MSE with the knowledge on the input spectrum in [13], we normalize $P[z]$ such as

$$P[e^{j\omega}] = P(\lambda\omega) \quad \text{for} \quad |\omega| \leq \omega_c \tag{17}$$

with $\omega_c = \pi/\lambda$.

Finally, we can state our main theorems:

Theorem 3. *Let the oversampling rate be λ and $\nu = \gamma + 1$ where γ is defined in Assumption 3. The MSE of the modulator is a function of ν and λ and is denoted as $D(\nu, \lambda)$. Then, $D(\nu, \lambda)$ satisfies*

$$D(\nu, \lambda) = \alpha_{opt} = D(\nu^\lambda, 1). \tag{18}$$

Since the uniform quantizer cannot outperform the modulator without oversampling, we have $D(\nu, 1) \leq ||P[z]||^2/\gamma$. It follows from (18) that

Theorem 4. *The MSE of the optimal modulator is upper bounded such that*

$$D(\nu, \lambda) \leq \left(\frac{1}{\nu^\lambda - 1}\right)||P[z]||^2. \tag{19}$$

Theorem 4 shows that the MSE of the Delta-Sigma modulator decays at the rate of $O(\nu^{-\lambda})$. On the other hand, the decay rate of the Delta-Sigma modulator having pre/post-filters designed with the knowledge of the input spectrum is $O(\nu^{-\lambda}/\lambda)$ [13, Theorem 6]. The decay rate is faster than the conventional Delta-Sigma modulator by a factor of $1/\lambda$, which is the benefit of availability of the input spectrum.

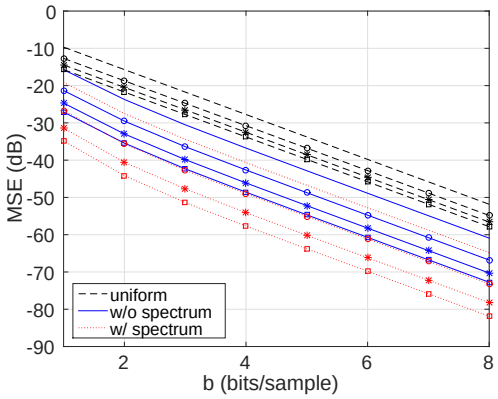

Fig. 3. MSEs of the optimal feedback quantizer, the optimal feedback quantizer [13] (dotted curve), and the uniform quantizer (dashed curve) with different oversampling rates λ, for a colored input, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

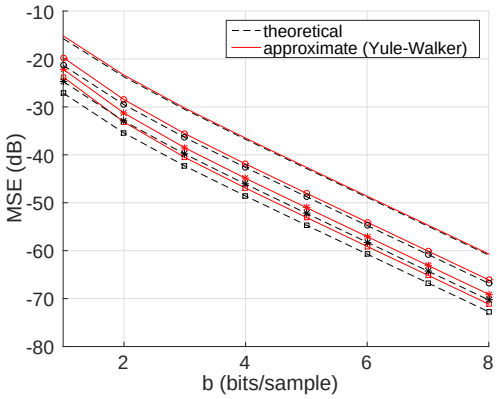

Fig. 4. MSEs of the feedback quantizers with ideal feedback filters and feedback quantizers with IIR feedback filters of order four approximated by the Yule-Walter method for different oversampling rates λ, where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively.

4. NUMERICAL EXAMPLES

To validate our analysis, we consider a continuous-time system of order four whose transfer function is

$$P(s) = \frac{1.029s^3 + 4.589s^2 + 7.146s + 3.882}{s^4 + 5.088s^3 + 9.789s^2 + 8.296s + 2.548}. \quad (20)$$

We model the continuous-time input signal as a stationary process with a zero mean and a spectrum given by $S(\omega) = c/|j\omega + 2.62|^2$, where c is a constant to normalize the sampled signal. We discretize these with a sampling period $T_s = 0.1$ to obtain the discrete-time system $P[\mathrm{z}]$ and the input x.

The loading factor is defined as $L_f = L/\sigma_u = 2^b d/(2\sigma_u)$ [17], which regulates the frequency of the overloading. We set it to be four. For $b = 1, 2, \ldots, 8$, we have $\gamma = 3 \cdot 2^{2b}/L_f^2$ [17]. Then, for a given γ, we numerically find the optimal α from (15) and (16) that is the minimum MSE replacing $p(\omega)$ by $p_\lambda(\omega)$ in (14).

For the oversampling ratio $\lambda = 1, 2, 3, 4$, Fig. 3 compares the MSEs of the optimal feedback quantizer, the optimal feedback quantizer with the pre-/post-filters [13] (dotted curve), and the uniform quantizer (dashed curve), where $\circ$, $*$, and $\square$ correspond to the oversampling ratios $\lambda = 2$, $\lambda = 3$, and $\lambda = 4$, respectively. For every quantizer, oversampling reduces the MSEs. However, we also find that oversampling is not so effective, since the number of bits per unit time is λb.

The feedback quantizer has an approximately 10 dB gain against the uniform quantizer that is enabled by utilizing the feedback filter that is optimized based on the system $P[\mathrm{z}]$. A further gain is obtained by exploiting the input spectrum for the quantizer having an optimized feedback filter and pre-/post-filters. For all quantizers, as the oversampling ratio increases, the MSE decreases and the increment of the MSE gain decreases.

In Fig. 3, we have utilized ideal feedback filters both for the feedback quantizer and the feedback quantizer with the pre-/post-filters, which cannot be implemented in practice. We approximate the ideal feedback filters for the optimal feedback quantizers using IIR filters of order four by the Yule-Walker method [18]. We just normalize the approximated filter so that the head of its impulse response is unity.

Fig. 4 illustrates the MSEs of the feedback quantizers with ideal optimal feedback filters and the feedback quantizers with feedback filters of order four approximated by the Yule-Walker method. The approximation by the Yule-Walker method suffers a small loss due to the error by the normalization.

5. CONCLUSIONS

We have presented the rate-distortion analysis of quantizers with error feedback. We have shown that the amplitude response of the optimal error feedback filter that minimizes the MSE can be parameterized by one parameter and can be found numerically. With the optimal error feedback filter, the relationship between the number of bits used for the quantization and the achievable MSE has been clarified. Numerical examples have been provided to demonstrate our analysis and synthesis.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP16K06356.

6. REFERENCES

[1] K. Sayood, *Introduction to data compression*. Newnes, 2012.

[2] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.

[3] W. Higgins and D. Munson, "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 6, pp. 963–973, Dec 1982.

[4] M. Nagahara and Y. Yamamoto, "Frequency domain min-max optimization of noise-shaping delta-sigma modulators," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2828–2839, June 2012.

[5] S. Callegari and F. Bizzarri, "Output filter aware optimization of the noise shaping properties of Δ Σ modulators via semi-definite programming," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 9, pp. 2352–2365, Sept 2013.

[6] S. Ohno, Y. Wakasa, and M. Nagata, "Optimal error feedback filters for uniform quantizers at remote sensors," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3866–3870.

[7] M. R. Tariq and S.Ohno, " Unified LMI based Design of $\Delta\Sigma$ Modulators, " *EURASIP Journal on Advances in Signal Processing*, 2016:29, 2016.

[8] C. Mullis and R. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 9, pp. 551–562, Sep 1976.

[9] T. Laakso and I. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Transactions on Signal Processing*, vol. 40, no. 5, pp. 1096–1107, May 1992.

[10] N. Thao, "Vector quantization analysis of $\Sigma\Delta$ modulation," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 808–817, Apr 1996.

[11] I. Daubechies and R. DeVore, "Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Annals of mathematics*, pp. 679–710, 2003.

[12] P. Deift, F. Krahmer, and C. S. Güntürk, "An optimal family of exponentially accurate one-bit sigma-delta quantization schemes," *Communications on Pure and Applied Mathematics*, vol. 64, no. 7, pp. 883–919, 2011.

[13] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3871–3890, Aug 2008.

[14] Tran-Thong and B. Liu, "Error spectrum shaping in narrow-band recursive filters," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 200–203, Apr 1977.

[15] J. Tuqan and P. Vaidyanathan, "Statistically optimum pre-and postfiltering in quantization," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 44, no. 12, pp. 1015–1031, 1997.

[16] S. Ohno, T. Shiraki, M. Rizwan Tariq, and M. Nagahara, "Rate-Distortion analysis of quantizers with error feedback," `arXiv:1609.01383 [cs.SY]`.

[17] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[18] B. Friedlander and B. Porat, "The modified Yule-Walker method of ARMA spectral estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-20, no. 2, pp. 158–173, March 1984.