# Analysis of the Quantization Error in Digital Multipliers with Small Wordlength

Günter Dehner
Ingenieurbüro Dehner
D-91058 Erlangen
Buckenhofer Weg 52
guenter.dehner@ib-dehner.de

Rudolf Rabenstein and Maxmilian Schäfer
Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg
D-91058 Erlangen, Cauerstr. 7
Rudolf.Rabenstein@FAU.de, Max.Schaefer@FAU.de

Christian Strobl
E-T-A
Elektrotechnische Apparate GmbH
D-90518 Altdorf, Industriestr. 8-10
Christian.Strobl@e-t-a.de

***Abstract*—The analysis of the quantization error in fixed-point arithmetic is usually based on simplifying assumptions. The quantization error is modelled as a random variable which is independent of the quantized variable. This contribution investigates the wordlength reduction of a digital multiplier in greater detail. The power spectrum of the quantization is expressed by the power spectrum of the multiplier input. The analytical results agree with measurements of the quantization error. The presented error model is shown to be superior to the simplified one for wordlengths in the range of eight bit.**

***Index Terms*—Finite wordlength effects, error analysis, quantization, digital arithmetic, Gaussian processes.**

## I. Introduction

Quantization effects in digital filters have been extensively studied when fixed-point arithmetic was mandatory for digital processing systems with real-time capability. The interest declined when floating-point processors with high clock rates became available. The classical knowledge on the different aspects of fixed-point quantization is found in standard textbooks e.g. [1]–[4].

Recently fixed-point arithmetic with short wordlength regains attention for reasons like cost and power consumption. Intelligent sensors and devices require front-end data processing and machine-learning while cost constraints in the centrange prevail. Other applications require long battery life such that power consumption has to be minimized by saving digits.

DSP systems have been designed by optimization using search algorithms in [5], [6]. Current research interests lie in the design and evaluation of systems with fixed-point arithmetic by both analytic methods and simulation [7], [8]. Descision errors in communication systems are studied in [9] where quantization noise is assumed to be signal independent. The estimation of the output power spectrum by a linear filter model is investigated in [10]. Filters with poles close to the unit circle are designed in [11] where the noise analysis assumes a uniform roundoff noise distribution. The implementation of linear filters is discussed in [12], roundoff errors are modelled by an additive system. Very recently, [13] consider $L_1$-norm error bounds for wave digital filters and [14] investigate signal quantization for control applications.

This contribution reconsiders the power spectrum of the quantization error of a digital multiplier. There are no assumptions either on the probability mass of the discrete error values nor on their correlation. Instead the autocorrelation of the quantization error is expressed in terms of the autocorrelation of the multiplier input. Here a joint normal distribution is assumed and the finite wordlength of the input is taken into account.

Secs. II and III introduce quantization and finite wordlength multiplication. After a review of a classical quantization model in Sec. IV, a more detailed multiplier model is presented in Sec. V. It allows to derive explicit relations of the power spectrum of the quantization error in Sec. VI. The analytical results are compared to measurements of the quantization error in Sec. VII.

## II. Quantization

### A. Quantization of continuous and discrete variables

The term quantization is used both for the conversion of an analog signal into a digital one (AD-conversion) and for the wordlength reduction in digital systems.

In analog-to-digital conversion quantization describes the mapping of a continuous quantity $x$ to another quantity $y$ which is restricted to a finite set of values. In binary representation with a wordlength of $w$ bit there are up to $2^w$ different states. Typical number representations like two's complement or sign-and-magnitude representation consist of a sign bit and a fractional part with $w-1$ bits. The smallest step change in quantized values is $Q = 2^{1-w}$ such that the number representation for $y$ is

$$y = \lambda Q \quad \text{with} \quad \lambda \in \mathbb{L}\,. \tag{1}$$

For sign-and-magnitude representation the set $\mathbb{L}$ is given by

$$\mathbb{L} = \{\lambda \in \mathbb{Z} | 1 - Q^{-1} \leq \lambda \leq Q^{-1} - 1\}\,, \tag{2}$$

while two's complement representation permits also the value $\lambda = -Q^{-1} = -2^{w-1}$.

In digital filters quantization denotes the mapping from a digital signal $x$ with wordlength $w_x$ to another digital signal $y$ with a smaller wordlength $w_y < w_x$. The description by (1) and (2) holds for both signals with sets $\mathbb{L}_x$ and $\mathbb{L}_y$ described by $Q_x = 2^{1-w_x}$ and $Q_y = 2^{1-w_y}$, respectively.

Funded by the Bavarian Research Foundation
(Bayerische Forschungsstiftung), grant number 1108-13.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

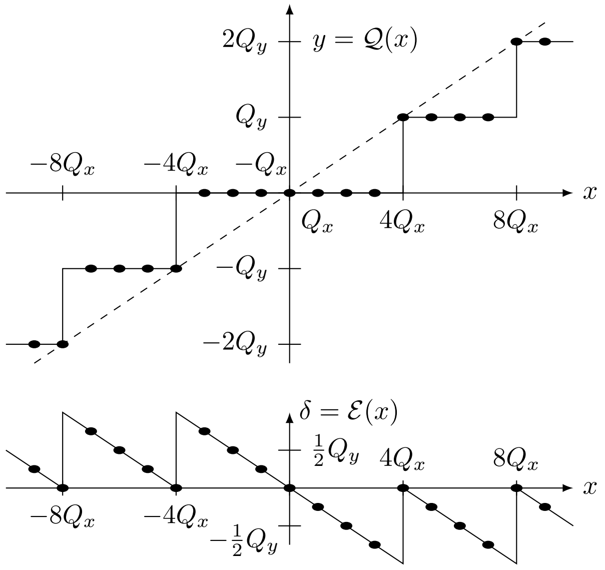

Fig. 1. Top: Quantization law $y = \mathcal{Q}(x)$ for sign-and-magnitude truncation. Solid line: quantization of continuous variable $x$, dots: quantization of discrete variable $x$ with quantization step $Q_x$, dashed line: identity $y = x$. Bottom: Quantization error law $\delta = \mathcal{E}(x) = \mathcal{Q}(x) - x$ for the above quantization law.

### *B. Quantization law and quantization error law*

The quantization process from Eq. (1) is described by a mapping through the quantization law $\mathcal{Q}(x)$ as $y = \mathcal{Q}(x)$ . Fig. 1 shows an example for wordlength reduction from $w_x$ to $w_y = w_x - 2$ by sign-and-magnitude truncation, as applied for the reduction to limit cycles. The corresponding quantization error $\delta$ is given by the difference

$$\delta(x) = y(x) - x = \mathcal{Q}(x) - x = \mathcal{E}(x) \ . \qquad (3)$$

Here $\mathcal{E}(x)$ denotes the quantization error law which maps the quantity $x$ onto its quantization error $\delta(x)$, see Fig. 1.

## III. Finite Wordlength Multiplication

The multiplication of finite wordlength signals is explained in Fig. 2. The input signal $v$ with wordlength $w_v$ is multiplied by the value of a multiplier coefficient $c$ with wordlength $w_c$. The value of the multiplication is $x$ and its correct representation requires a wordlength of $w_x = w_v + w_c - 1$. Quantization according to the quantization law $\mathcal{Q}(x)$ reduces the wordlength of the resulting signal $y$ to $w_y$. A typical value is $w_y = w_v$.

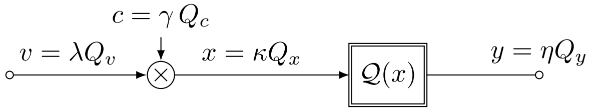

Fig. 2. Finite wordlength multiplication. Input signal $v$, multiplier coefficient $c$, exact multiplication value $x$, quantized multiplication value $y$, quantization law $\mathcal{Q}(x)$. The integers $\lambda$, $\gamma$, $\kappa$, $\eta$ and the various quantization steps $Q$ are explained in Secs. II-A and V-B.

In digital filters, the variables $v$, $x$, $y$ are sample values of discrete-time signals, with time variable $k$, i.e. $v(k)$, $x(k)$, $y(k)$. Also the multiplier coefficient can vary with time.

Thus also the quantization error $\delta(k) = y(k) - x(k)$ varies in each time step. It depends in a deterministic way on the current input signal, the multiplier value and the quantization law. However, this dependency is not easy to quantify. Furthermore the input signal itself might be a random signal.

Therefore the quantization error is usually modelled as a random variable. Its properties depend on certain assumptions as discussed in the following sections.

## IV. The $Q^2/12$ Quantization Model

The most simple and most widely used quantization model is based on the assumption that the quantization error is uniformly distributed (e.g. [1]–[4]). For quantization by rounding the variance of the quantization error is calculated as

$$\sigma_\delta^2 = \int_{-\infty}^{\infty} \delta^2 p_\delta(\delta)\, d\delta = \frac{1}{Q_y}\int_{-Q_y/2}^{Q_y/2} \delta^2\, d\delta = \frac{Q_y^2}{12} \ . \qquad (4)$$

The simplicity of this result is based on the assumption that the quantization error has a probability density function $p_\delta(\delta)$, which is independent of the density of the quantized signal.

When considering the quantization error $\delta(k)$ as a random signal, a further assumption on the covariance of subsequent samples is required. It is commonly assumed, that these samples are independent of the multiplier value, not correlated with the unquantized variable and uncorrelated among themselves. The quantization error is then a white noise sequence with a constant power spectral density.

The above assumptions lead to a simple and popular model for the quantization noise. Its effect is described as an additive white noise source with variance $Q^2/12$. Including additive noise sources at all quantization points of a digital system allows to carry out the analysis of roundoff noise in the familiar framework of linear and time-invariant systems. Further details can be found e.g. in [1]–[4].

This additive noise source model has proven to be successful for the analysis of fixed point arithmetic when the wordlength after quantization is not too small (e.g. above 8 bit). However, deviations from the white noise source can be observed for systems with smaller wordlength and for certain multiplier values (see Sec. VII). Thus a more detailed analysis of the sequence of quantization errrors is required.

## V. A detailed digital multiplier model

An in-depth analysis of the quantization error of a digital multiplier is performed here in three steps:

- Model the input signal $v(k)$ as a random process.
- Derive the properties of the exact multiplier output $x(k)$.
- Derive the properties of the quantization error $\delta(k)$.

### *A. Multiplier input signal $v(k)$*

The multiplier input signal $v(k)$ is assumed to be a realization of a stationary random process with the joint probability mass function

$$f_{v_1,v_2}(v_1,v_2;m) = \sum_{\lambda_1\in\mathbb{L}_v}\sum_{\lambda_2\in\mathbb{L}_v} p_{\lambda_1,\lambda_2}\delta(v_1-\lambda_1 Q_v)\delta(v_2-\lambda_2 Q_v). \qquad (5)$$

The variables $v_1$ and $v_2$ are values spaced $m$ samples apart, i.e. $v_1 = v(k)$ and $v_2 = v(k+m)$. The set of integer values $\mathbb{L}_v$ is defined as in (2) with quantization step $Q_v$.

The probabilities $p_{\lambda_1,\lambda_2}$ are calculated from a two-dimensional normal distribution $\mathcal{N}_{vv}$ as

$$p_{\lambda_1,\lambda_2} = \int_{\Omega_{\lambda_1}} \int_{\Omega_{\lambda_2}} \mathcal{N}_{vv}(v_1, v_2|\mathbf{0}, \boldsymbol{R})\, dv_1\, dv_2, \tag{6}$$

with the joint probability density function

$$\mathcal{N}_{vv}(v_1, v_2|\mathbf{0}, \boldsymbol{R}) = \frac{1}{2\pi\sigma_v^2\sqrt{|\boldsymbol{R}|}} \exp\left(-\frac{1}{2\sigma_v^2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{v}\right), \tag{7}$$

defined by the vector $\boldsymbol{v}$ and the covariance matrix $\boldsymbol{R}$

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}. \tag{8}$$

The covariance matrix $\boldsymbol{R}$ contains the normalized autocorrelation $r$ which is related to the autocorrelation sequence $R_{vv}(m)$

$$r = r(m) = \frac{1}{\sigma_v^2} R_{vv}(m). \tag{9}$$

The integration regions $\Omega_\lambda$ for $\lambda = \lambda_1$ and $\lambda = \lambda_2$ in (6) are centered around the quantized values $v = \lambda Q_v$ and also include the tails of the distribution

$$\Omega_\lambda = \begin{cases} [-\infty, \frac{1}{2}Q_v - 1] & \lambda = 1 - Q_v^{-1} \\ [(\lambda - \frac{1}{2})Q_v, (\lambda + \frac{1}{2})Q_v] & 1 - Q_v^{-1} < \lambda < Q_v^{-1} - 1 \\ [-\frac{1}{2}Q_v + 1, \infty] & \lambda = Q_v^{-1} - 1\,. \end{cases} \tag{10}$$

### B. Exact multiplier output signal $x(k)$

The corresponding relations for the variable $x$ after exact multiplication with a coefficient $c$ follow in a straightforward way from Sec. V-A. The discrete values of $x$ are exact multiplies of $v$ as shown in Fig. 2. Thus also the standard deviation $\sigma_v$ of the joint probability density (7) is scaled accordingly. The resulting relations are compiled as

$$x = cv = \lambda\gamma\, Q_v Q_c = \kappa\, Q_x, \qquad \sigma_x = c\,\sigma_v, \tag{11}$$

with $Q_x = Q_v Q_c$ and $\kappa = \lambda\gamma$.

The underlying joint probability density becomes

$$\mathcal{N}_{xx}(x_1, x_2|\mathbf{0}, \boldsymbol{R}) = \frac{1}{2\pi\sigma_x^2\sqrt{|\boldsymbol{R}|}} \exp\left(-\frac{1}{2\sigma_x^2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{x}\right). \tag{12}$$

Integration around the discrete values $\kappa Q_x$ gives the joint probabilities $p_{\kappa_1,\kappa_2}$. The integration regions $\Omega_{\kappa_1}$ and $\Omega_{\kappa_2}$ are defined similar to (10) with $Q = Q_x$. With

$$p_{\kappa_1,\kappa_2} = \int_{\Omega_{\kappa_1}} \int_{\Omega_{\kappa_2}} \mathcal{N}_{xx}(x_1, x_2|\mathbf{0}, \boldsymbol{R})\, dx_1\, dx_2 \tag{13}$$

follows the joint probability mass function

$$f_{x_1,x_2}(x_1, x_2; m) = \sum_{\kappa_1 \in \mathbb{L}_x} \sum_{\kappa_2 \in \mathbb{L}_x} p_{\kappa_1,\kappa_2} \delta(x_1 - \kappa_1 Q_x)\delta(x_2 - \kappa_2 Q_x), \tag{14}$$

which is required for the calculation of the quantization error. The index set $\mathbb{L}_x$ is defined as in (2) with $\kappa$ and $Q_x$.

### C. Quantization error $\delta(k)$

The quantization error depends on the quantization law $\mathcal{Q}(x)$ or directly on the quantization error $\mathcal{E}(x)$, see (3) and Fig. 3. The notation in the figure emphasizes that the input signal $x(k)$ represented by the index $\kappa(k)$ and the quantization error $\delta(k)$ adopt different values in each time step $k$.

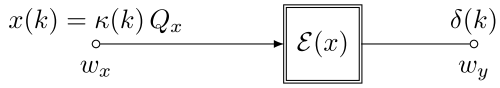

Fig. 3. Description of a quantizer which reduces the wordlength $w_x$ after multiplication to $w_y$. The quantization error sequence $\delta(k)$ is determined by the nonlinear quantization error law $\mathcal{E}(x)$.

Of interest is the noise power of the quantization error or – more general – its autocorrelation function $R_{\delta\delta}(m)$. It can be obtained from the joint probability mass function in (14) as

$$\begin{aligned} R_{\delta\delta}(m) &= \iint\limits_{-\infty}^{\infty} \mathcal{E}(x_1)\mathcal{E}(x_2) f_{x_1,x_2}(x_1, x_2; m)\, dx_1\, dx_2 \\ &= \sum_{\kappa_1 \in \mathbb{L}_x} \sum_{\kappa_2 \in \mathbb{L}_x} p_{\kappa_1,\kappa_2}\, \mathcal{E}(\kappa_1 Q_x)\, \mathcal{E}(\kappa_2 Q_x)\,. \end{aligned} \tag{15}$$

The probabilities $p_{\kappa_1,\kappa_2}$ depend via (8), (9) and (12) on the autocorrelation sequence $R_{vv}(m)$ at the input of the multiplier. Thus (15) provides an expression for the dependency of the autocorrelation sequence $R_{\delta\delta}(m)$ of the quantization error on the autocorrelation sequence $R_{vv}(m)$ at the input.

It would be even more useful to have a corresponding relation in the frequency domain, i.e. to express the power spectrum of the quantization error by the power spectrum at the input.

## VI. Power Spectrum of the Quantization Error

Power spectra in nonlinear control circuits have been investigated in [15]. A more direct relation than the one suggested by (15) can be found through a representation of the joint normal density (12) by Mehler's formula [15, Ch. XVII 2.1], [16], [17]. This representation has been introduced to the analysis of quantization errors in digital filters by Meyer [18] based on previous work [19].

### A. Series expansion of the autocorrelation sequence

Mehler's formula expands the joint probability density function (12) into the product of Hermite polynomials

$$\begin{aligned} &\mathcal{N}_{xx}(x_1, x_2|\mathbf{0}, \boldsymbol{R}) = \\ &\frac{1}{2\pi\sigma_x^2} \exp\left(-\tfrac{1}{2\sigma_x^2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right) \sum_{n=0}^{\infty} \frac{r^n(m)}{n!} H_n(\tfrac{x_1}{\sigma_x}) H_n(\tfrac{x_2}{\sigma_x})\,. \end{aligned} \tag{16}$$

An outline of the proof is given in the appendix. Note that unlike in (12) the dependence on the normalized autocorrelation $r(m)$ is of polynomial form which makes the subsequent analysis much simpler.

Following the approach from [18] insert (16) into (13) to obtain a series expansion of the probabilities $p_{\kappa_1,\kappa_2}$

$$p_{\kappa_1,\kappa_2} = \frac{1}{2\pi\sigma_x^2} \sum_{n=0}^{\infty} \frac{r(m)^n}{n!}\, a_n(\kappa_1)\, a_n(\kappa_2)\;, \qquad (17)$$

where the coefficients $a_n(\kappa)$ for $\kappa = \kappa_{1/2}$ depend on the region of integration $\Omega_\kappa$

$$a_n(\kappa) = \int_{\Omega_\kappa} \exp\left(-\frac{x^2}{2\sigma_x^2}\right) H_n\left(\frac{x}{\sigma_x}\right) dx\;. \qquad (18)$$

Now inserting (17) into (15) expresses the autocorrelation sequence of the quantization error by the normalized autocorrelation sequence $r(m)$ as

$$R_{\delta\delta}(m) = \frac{1}{\sigma_v^2} \sum_{n=0}^{\infty} b_n^2\; r^n(m), \qquad (19)$$

with

$$b_n = \frac{1}{c}\frac{1}{\sqrt{2\pi\, n!}} \sum_{\kappa\in\mathbb{L}_x} a_n(\kappa)\, \mathcal{E}(\kappa Q_x)\;. \qquad (20)$$

Note that the coefficients $a_n(\kappa)$ depend only on the assumed Gaussian distribution and the index $\kappa$ of the quantized value, while the coefficients $b_n$ include the quantization error law $\mathcal{E}(x)$ and the multiplier coefficient $c$.

Finally the normalized autocorrelation $r(m)$ can be converted into the autocorrelation sequence at the input by (9)

$$R_{\delta\delta}(m) = \sum_{n=0}^{\infty} \frac{b_n^2}{\sigma_v^{2(n+1)}}\, R_{vv}^n(m)\;. \qquad (21)$$

### *B. Power spectrum*

The discrete-time Fourier transform turns the autocorrelation sequence $r(m)$ into the power spectrum $S(e^{j\Omega})$. The square $r^2(m)$ turns into a frequency domain circular convolution of the power spectrum with itself and higher powers are turned into multiple convolutions

$$S_{vv}(e^{j\Omega}) = \sum_{m=-\infty}^{\infty} R_{vv}(m) e^{-jm\Omega}\;, \qquad (22)$$

$$\frac{1}{2\pi} S_{vv}(e^{j\Omega}) * S_{vv}(e^{j\Omega}) = \sum_{m=-\infty}^{\infty} R_{vv}^2(m) e^{-jm\Omega}\;. \qquad (23)$$

Thus the power spectrum of the quantization error $S_{\delta\delta}(e^{j\Omega})$ can be expressed by the power spectrum of the multiplier input $S_{vv}(e^{j\Omega})$ (the symbol $\overset{n}{*}$ denotes $n$-fold circular convolution)

$$S_{\delta\delta}(e^{j\Omega}) = \frac{b_0^2}{\sigma_v^2}\delta(\Omega) + \frac{b_1^2}{\sigma_v^4} S_{vv}(e^{j\Omega}) + \\ + \sum_{n=2}^{\infty} \frac{b_n^2}{\sigma_v^{2(n+1)}} \frac{1}{(2\pi)^{n-1}}\, S_{vv}(e^{j\Omega}) \overset{n-1}{*} S_{vv}(e^{j\Omega})\;. \qquad (24)$$

The power spectrum $S_{\delta\delta}(e^{j\Omega})$ consists of

- a DC-term representing a bias in the quantized signal,
- a term which corresponds to the power spectrum $S_{vv}(e^{j\Omega})$,
- all further terms with multiple convolutions of the input power spectrum.

The relative weighting of these different contributions is determined by the properties of the quantization law.

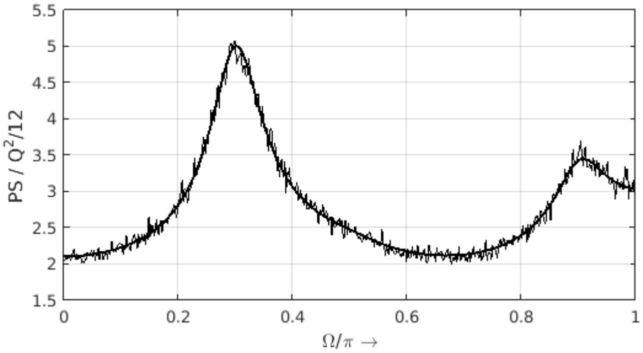

Fig. 4. Power spectrum of the quantization error as described in Sec. VII. Solid line: Analytical result from the series (24). Wiggly line: Measurement by simulation of the system from Fig. 2

## VII. Numerical Results

As an example consider a white Gaussian noise signal with variance $\sigma = 0.25$ filtered through a discrete-time system with transfer function

$$F(z) = \frac{a(z-1)}{(z-z_\infty)(z-z_\infty^*)}, \quad \text{with} \quad z_\infty = 0.95\, e^{j0.3\pi}, \qquad (25)$$

such that the power spectrum at the multiplier input is

$$S_{vv}(e^{j\Omega}) = |F(e^{j\Omega})|^2\;. \qquad (26)$$

The pre-filter $F(z)$ is scaled by a parameter $a$ in (25) such that $\sigma_v = \sigma = 0.25$. Quantizing the pre-filter ouput to a wordlength $w_v$ gives the signal $v(k)$. The discrete-time variables are represented by sign and magnitude with wordlength $w_v = w_c = w_y = 8\,\text{bit}$. The exact result of the multiplication $x(k) = c\,v(k)$ with $c = 63\,Q_c = 63/128$ is truncated to $y(k)$ with wordlength $w_y$.

The power spectrum $S_{\delta\delta}(e^{j\Omega})$ of the quantization error was

- calculated analytically with (24),
- measured by the method from [2, Sec. 5.5.5] [20] ($N = 512$ samples of the spectrum, length of the test signal $2N$, average over $L = 1024$ measurements).

Both results are shown in Fig. 4 where the power spectrum $S_{\delta\delta}(e^{j\Omega})$ is scaled by $Q_y^2/12$. It is obvious that the power spectrum of the quantization error is stronlgy determined by the spectral shape of the multiplier input. The peak at $\Omega = 0.3\pi$ is directly related to the poles of the filter $F(z)$. A third harmonic appears at $\Omega = 0.9\pi$ caused by the nonlinearity of the quantization.

The measured power spectrum is well approximated by the series representation (24). Due to the form of the quantization error law in this case, only odd values $n$ appear in the series. The spectral shape is determined by the term with $n = 1$ and further by the convolution terms $n = 3, 5, \ldots, 13$. Further terms for $n > 13$ are flattened by the repeated convolutions and can be well approximated by a white spectrum. This flattening effect has already been described in [18].

## VIII. Conclusion

The noise power spectrum of the quantization error at the output of a finite-wordlength digital multiplier has been investigated. No assumptions on the properties of the noise power spectrum were made. Instead, the power spectrum at the output has been derived from the assumed properties of the signal power spectrum at the multiplier input. The mathematical analysis is based on an expansion of the Gaussian distribution into Hermite polynomials. The obtained analytical result compares favourably with measurements at a simulated finite-wordlength multiplier. The presented results provide a building block for the analysis of finite-wordlength implementations of digital filters and digital controllers.

## Appendix

Mehler's formula is a series expansion of a two-dimensional Gaussian function into the product of Hermite polynomials [16]. It is frequently quoted [15], [17], [18], [21]–[23] but the appearance of Mehler's formula varies, because different definitions and notations of the Hermite polynomials exist [22]. Since proofs are hard to find in the technical literature, a derivation of Mehler's formula is given here. It is based on an approach from the original paper [16] and uses results of [24].

A random variable with standard normal distribution has the probability density function $\mathcal{N}(x|0,1)$ and the characteristic function $\Phi_x(\omega) = \exp(-\frac{1}{2}\omega^2)$ [24, Table 5.2]

$$\mathcal{N}(x|0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_x(\omega) e^{-j\omega x}\, d\omega \;, \quad (27)$$

such that the exponential in $\mathcal{N}(x|0,1)$ can be expressed as

$$e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{\omega^2}{2} + j\omega x)}\, d\omega \;. \quad (28)$$

Now consider the Hermite polynomials e.g. from [24, Sec. 7.4]

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}} \;. \quad (29)$$

With (28) and repeated differentiation inside the integral follows

$$e^{-\frac{x^2}{2}} H_n(x) = \frac{1}{\sqrt{2\pi}} j^n \int_{-\infty}^{\infty} \omega^n e^{-(\frac{\omega^2}{2} + j\omega x)}\, d\omega \;. \quad (30)$$

The joint normal density function $\mathcal{N}_2(x_1, x_2|\mathbf{0}, \boldsymbol{R})$ from (16) has the joint characteristic function [24, Sec. 6.5]

$$\begin{aligned} \Phi_{x_1 x_2}(\omega_1, \omega_2) &= \exp\left(-\tfrac{1}{2}(\omega_1^2 + \omega_2^2 + 2r\omega_1\omega_2)\right) \quad (31) \\ &= \exp\left(-\tfrac{1}{2}(\omega_1^2 + \omega_2^2)\right) \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} r^n \omega_1^n \omega_2^n \;, \end{aligned}$$

such that it can be expressed as

$$\mathcal{N}_2(x_1, x_2) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} \Phi_{x_1 x_2}(\omega_1, \omega_2) e^{-j(\omega_1 x_1 + \omega_2 x_2)}\, d\omega_1 d\omega_2 \;.$$

Inserting $\Phi_{x_1 x_2}(\omega_1, \omega_2)$ in the form of (31), interchanging summation and integration and replacing each integral by (30) for ($x_1$, $\omega_1$) and for ($x_2$, $\omega_2$) gives (16) for $\sigma_x = 1$.

## References

[1] L. B. Jackson, *Digital Filters and Signal Processing*. Boston, USA: Kluwer Academic Publishers, 1995.

[2] H. W. Schüßler, *Digitale Signalverarbeitung*, 5th ed. Berlin Heidelberg: Springer-Verlag, 2008, vol. 1.

[3] S. Mitra, *Digital Signal Processing*. Boston, USA: McGraw-Hill, 2011.

[4] J. Proakis and D. Manolakis, *Digital Signal Processing*. London, UK: Pearson Education, 2013, vol. 4.

[5] W. Sung and K.-I. Kum, "Simulation-based word-length optimization method for fixed-point digital signal processing systems," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 3087–3090, Dec 1995.

[6] K.-I. Kum and W. Sung, "Combined word-length optimization and high-level synthesis of digital signal processing systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 20, no. 8, pp. 921–930, Aug 2001.

[7] K. Parashar, D. Menard, and O. Sentieys, "Accelerated performance evaluation of fixed-point systems with un-smooth operations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 4, pp. 599–612, April 2014.

[8] R. Nehmeh, D. Menard, A. Banciu, T. Michel, and R. Rocher, "Integer word-length optimization for fixed-point systems," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 8321–8325.

[9] K. Parashar, R. Rocher, D. Menard, and O. Sentieys, "Analytical approach for analyzing quantization noise effects on decision operators," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 1554–1557.

[10] K. Parashar, D. Menard, R. Rocher, and O. Sentieys, "Estimating frequency characteristics of quantization noise for performance evaluation of fixed point systems," in *18th European Signal Processing Conference (EUSIPCO)*, Aug 2010, pp. 552–556.

[11] L. Harnefors, "Implementation of resonant controllers and filters in fixed-point arithmetic," *IEEE Trans. Ind. Electron.*, vol. 56, no. 4, pp. 1273–1281, April 2009.

[12] T. Hilaire and B. Lopez, "Reliable implementation of linear filters with fixed-point arithmetic," in *IEEE Workshop on Signal Processing Systems (SiPS)*, Oct 2013, pp. 401–406.

[13] A. Volkova and T. Hilaire, "Fixed-point implementation of lattice wave digital filter: Comparison and error analysis," in *23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1118–1122.

[14] M. Takeya, Y. Kawamura, and S. Katsura, "Data reduction design based on delta-sigma modulator in quantized scaling-bilateral control for realizing of haptic broadcasting," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1962–1971, March 2016.

[15] H. Schlitt, *Stochastische Vorgänge in linearen und nichtlinearen Regelkreisen*. Braunschweig, Germany: Friedr. Vieweg & Sohn, 1968.

[16] F. G. Mehler, "Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung." *J. Reine Angew. Math.*, vol. 66, pp. 161–176, 1866.

[17] W. F. Kibble, "An extension of a theorem of Mehler's on Hermite polynomials," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 41, pp. 12–15, 6 1945. [Online]. Available: http://journals.cambridge.org/article_S0305004100022313

[18] R. Meyer, "Zur Realisierung von digitalen Systemen mit Festkomma-Signalprozessoren," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, 1994.

[19] G. Dehner, "Ein Beitrag zum rechnergestützten Entwurf rekursiver digitaler Filter minmalen Aufwands," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, 1976.

[20] H. Schüßler and Y. Dong, "A new method for measuring the performance of weakly nonlinear systems," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1989, pp. 2089–2092.

[21] G. Szegö, *Orthogonal Polynomials*, 4th ed., ser. Colloquium Publications. Providence, Rhode Island, USA: American Mathematical Society, 1985, vol. 23, reprint.

[22] "Mehler kernel," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Mehler_kernel

[23] E. W. Weisstein, "Mehler's Hermite polynomial formula," From MathWorld–A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/MehlersHermitePolynomialFormula.html

[24] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. Boston, USA: WCB/McGraw-Hill, 2002.