Spectral Density Shaping of Quantisation Error Using Dithering

Ahmad Faza, John Leth, and Arnfinn A. Eielsen

***Abstract*— A method for shaping the power spectral density (PSD) of the total error due to uniform quantisation is proposed. It utilises non-subtractive dithering, generated with a joint specification of the probability density function (PDF) and the PSD. If the dither PDF has certain properties, the expected value of the quantiser output can be linearised and the variance of the error (the error power) can be made independent of the quantiser input. It is demonstrated that making the error power independent of the input enables shaping the error PSD as a separate problem. The method relies on the use of a linear noise colouring filter and a static non-linear transform, which imposes some restrictions on the specification of the error PSD. However, it is possible to synthesise a range of filters that can be used to shape the PSD of the error. Simulations are provided to verify and illustrate the operation of the method.**

I. INTRODUCTION

Quantisation and re-quantisation are fundamental operations in digital signal processing, digital-analogue conversion, power electronics and measurement systems. Error is introduced since only a discrete subset of values can be represented [1,2]. Several methods can be used to shape the power spectral density (PSD) of quantisation error. Combined with oversampling, shaping the error PSD can be used to reduce the effective error due to quantisation by additional filtering in the frequency domain where the error power is concentrated. $\Delta\Sigma$-modulation uses quantiser model feedback to shape the PSD at the output [3]. As the quantiser is discontinuous, it becomes a chaotic system with an input-dependent, empirical sense of stability [4]. An alternative is to use model predictive control (MPC) which can offer both higher performance and rigorous stability results [5,6], but comes with high computational cost. Learning control (LC) has been shown to significantly suppress quantisation error [7], and can use actual output measurements, not only model feedback, but is restricted to periodic input signals.

Dithering is a feed-forward method with similar properties. An external signal is added to the input of the quantiser, and the resulting quantiser error can have a spectral distribution [8,9]. Dithering has been shown to mitigate the effect of static non-linearities such as element mismatch in digital-analogue converters [10] as well as dynamic non-linearities [11,12] such as glitches in digital-analogue converters [13]–[15]. In this paper we study the use of non-subtractive dither (NSD). Alternatively, subtractive dithering (SD) [16] can be used, and may have more conducive properties. However it can only be utilised when the exact dither signal is available for subtraction after quantisation. This is challenging to achieve in practical systems [16], and severely limits the applicability of SD in many systems.

If the dither probability density function (PDF) is triangular with a range that is an integer multiple of the quantisation step, the averaged response of a quantiser can be linearised and the variance of the error, or for a zero-mean signal, the error power, can be made independent of the quantiser input [8]. For the special case of a triangular PDF (TPDF), it can be produced by applying a simple difference operation to a noise signal with rectangular PDF (RPDF), causing convolution of two RPDFs, and simultaneously producing a first-order high-pass PSD [8]. The shape of the dither PSD is in this case maintained in the output, but the noise-shaping effect is modest. Here we investigate the possibility of improving the error PSD shaping when dithering.

A dither signal can be produced using a pseudo-random number generator (PRNG) [17]. These typically produce a RPDF, but arbitrary PDFs can be realised e.g. by applying the inverse of the cumulative distribution function (CDF) to the PRNG output [17]. Coloured noise can be produced by spectral factorisation of a desired PSD [18], producing an implementable linear filter. Combining spectral factorisation with the inverse CDF method is not straight-forward, as the output of linear filters driven with noise will tend to a normal distribution, due to the central limit theorem [19], and conversely static non-linearities will tend to whiten the PSD of the input signal [20]. One solution to the problem of realising a noise-like signal with jointly specified PDF and PSD is to compensate for the whitening effect of the non-linearity when synthesising the colouring filter [21,22].

A. Contributions

Results from [8] and the method from [22] are used to provide a novel solution to the problem of shaping the PSD of the total error due to uniform quantisation when applying a non-subtractive dither signal. As TPDF dither with proper amplitude decouples the error power from the input to the quantiser, the method in [22] is extended to find the non-linear transformation necessary such that a desired PSD for the total error is produced in the output. Simulations show that a wide range of PSDs can be realised, and significantly improved ability to shape the error can be achieved compared to the PSDs realisable using the method in [8].

B. Notation

A definition is denoted by $\triangleq$ and the Fourier operator is denoted $\mathcal{F}$. All probabilistic considerations will be with respect to a fixed probability space $(\Omega, \mathbf{F}, \mathbf{P})$, and for a random variable, g, we typically omit the dependency on $\varpi \in \Omega$, we write $g = g(\varpi)$. The probability density

Ahmad Faza' (ahmad.faza@uis.no), Arnfinn A. Eielsen (eielsen@ux.uis.no) are with the Dept. of Energy and Petroleum Engineering, University of Stavanger, Norway

J. Leth (jjl@es.aau.dk) is with the Dept. of Electronic Systems, Automation & Control, Aalborg University, Denmark

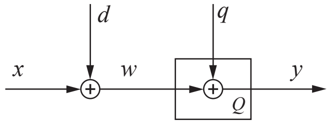

Fig. 1. Non-subtractively dithered quantiser, with input x, dither d, quantiser input w, output y, and quantisation error q.

function (PDF) of g is denoted by $p_g(g)$ and the characteristic function (cf) by $P_g(u) \triangleq \mathcal{F}[p_g](u) = E[e^{jug}]$, with the ith moment of g given by $E[g^i] \triangleq \int_{\mathbb{R}} g^i p_g(g)\,\mathrm{d}g$, and the cumulative distribution function (CDF) is denoted $F_g(a) \triangleq \int_{-\infty}^{a} p_g(u)\,\mathrm{d}u$. The notation is also used if g is a stochastic process, $g = g(t) = g(t, \varpi)$, i.e. all quantities are time dependent. If $\psi(\tau) \triangleq E(g(t)g(t+\tau))$ denotes the auto-covariance function (ACF) of g with delay operator τ, then the PSD $\Psi(\omega)$ of g is the Fourier transform $\Psi(\omega) = \mathcal{F}[\psi](\omega)$ with ω the angular frequency. The Heaviside step-function is defined as

$$\Gamma(u) \triangleq \begin{cases} 0\,, & u \le 0 \\ 1\,, & u > 0 \end{cases}\,,$$

and convolution is defined as

$$(f * h)(t) \triangleq \int_{-\infty}^{\infty} f(\tau)h(t-\tau)\,\mathrm{d}\tau\,.$$

II. ANALYTICAL FRAMEWORK

A. Uniform Quantisation

A mid-tread uniform quantiser, Q in Fig. 1, with step-size $\Delta \in \mathbb{R}_{>0}$ can be defined as $Q(w) \triangleq \Delta \left\lfloor \frac{w}{\Delta} + \frac{1}{2} \right\rfloor$ where $\lfloor \cdot \rfloor$ denotes the floor operator. The quantisation error $q(w)$ given an input w is defined as the function

$$q(w) \triangleq Q(w) - w\,. \tag{1}$$

The output can then be modelled as:

$$y = w + q(w) = Q(w)\,. \tag{2}$$

If we consider a mid-rise multi-bit quantiser with a word-size of B bits, it has 2^B output levels. With an output range between V_{min} and V_{max} the step-size is

$$\Delta = \frac{V_{\text{max}} - V_{\text{min}}}{2^B - 1}\,, \tag{3}$$

and in this case $Q(w)$ can be expressed as

$$Q(w) = \Delta \sum_{i=1}^{N_T} \left(\Gamma(w - T_i) - \frac{1}{2} \right)\,, \tag{4}$$

where $N_T = 2^B - 1$ is the number of quantisation levels, and the step-functions have the thresholds $T_i : i \in \{1, 2, 3, ...N_T\}$; where, $T_j = (j-i)\Delta + T_i$ for $j > i$ and $T_{2^{B-1}} = (V_{\text{max}} + V_{\text{min}})/2$.

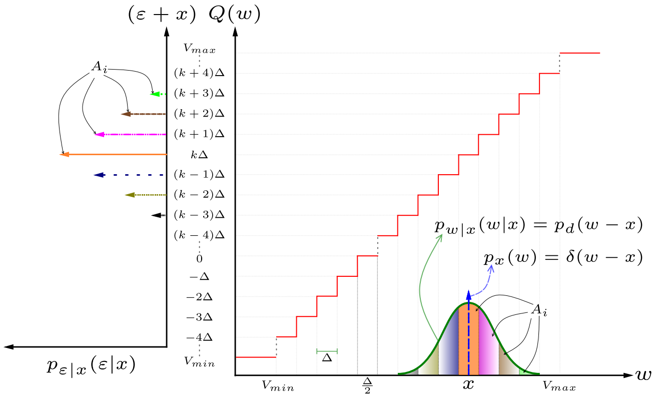

Fig. 2. NSD for a mid-tread uniform quantiser by dither d of PDF p_d. The conditional PDF (cPDF) $p_{\varepsilon|x}(\varepsilon|x)$ of the total error ε for a fixed input value x is an area sampled version of the quantiser input cpdf $p_{w|x}(w|x)$. A_i is the sampled area under $p_d(w-x)$ for $w \in [-\frac{\Delta}{2}+(k+i)\Delta, \frac{\Delta}{2}+(k+i)\Delta]$.

B. Non-subtractive Dithering (NSD)

Consider the quantiser configuration in Fig. 1. The total error ε is defined as the difference between the output y and input x, $\varepsilon \triangleq y - x$, to distinguish it from the quantisation error q in (1). For NSD $\varepsilon = Q(x+d) - x = q(x+d) + d$. From [8], the conditional PDF of ε given the input x is

$$p_{\varepsilon|x}(\varepsilon|x) = p_{-\infty} + p_{\infty} + \sum_{k=2}^{N_T - 1} p_k \tag{5}$$

with

$$\begin{aligned} p_k &= \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_d(w-x)\,\mathrm{d}w \\ p_{-\infty} &= \delta(\varepsilon + x - \Delta) \int_{-\infty}^{\frac{\Delta}{2}+\Delta} p_d(w-x)\,\mathrm{d}w \\ p_{\infty} &= \delta(\varepsilon + x - N_T\Delta) \int_{-\frac{\Delta}{2}+N_T\Delta}^{\infty} p_d(w-x)\,\mathrm{d}w \end{aligned}$$

where δ is the Dirac delta function, see Fig. 2. Note that $p_{\varepsilon|x}(\varepsilon|x)$ cannot be rendered independent of x by any choice of dither.

For given specifications of the PSD $S(\omega)$ of the total error ε we will in the sequel investigate how to generate a dither d which both linearises the uniform quantiser in mean, that is, $E[y]$ becomes a linear function of x, and induces a PSD $S(\omega)$ having the required specifications.

Regarding the linearisation, first note that $E[y] = x + E[\varepsilon]$. Now from [8] we know that $E[\varepsilon] = E[d]$ and $E[\varepsilon^2] = E[d^2] + \frac{\Delta^2}{12}$ whenever

$$G_d^{(m)}\left(\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbb{Z} - \{0\},\ m = 1, 2 \tag{6}$$

with $G_d(u) = \operatorname{sinc}(u)P_d(u)$, $G_d^{(m)}$ the mth derivative of G_d, and $\operatorname{sinc}(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u}$. Choosing a zero mean dither with a triangular PDF (and cf $P_d = \operatorname{sinc}^2(u)$) will result in (6) being fulfilled, see [8]. With this choice of dither it follows that $E[\varepsilon] = 0$ thus linearising the uniform quantiser

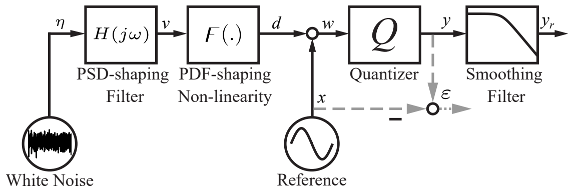

Fig. 3. Block diagram for the proposed method: White Gaussian process η, spectrally shaped Gaussian process v, dither with desired PDF d, input x, quantiser input w, quantiser output y, total error ε, and reconstructed output y_r.

in mean. Moreover, note how the total error ε is now a zero mean stationary process with $E[\varepsilon^2] = \Delta^2/4$. Hence, $\text{Var}(y) = E[\varepsilon^2]$ is constant and independent of the value of x. Furthermore, it follows from the choice of a TPDF dither, that the PSD S_y of the output y is just the sum of the PSD S of the total error ε and the PSD S_x of the input x. Thus shaping S also shapes S_y.

When it comes to spectrally shaping S, it is ideally desirable to shape S such that it has minimal power content at frequencies where S_x has significant power content. For example, the reconstructed output y_r after the smoothing filter in Fig. 3 contains the low-pass power content of y where S_x is concentrated; therefore, it is advantageous to shape S to have a high-pass power content away from S_x such that $y_r \approx x$. The best shaping performance known to be achievable using TPDF NSD is in (Sec.II-C & Sec.III-E [8]):

$$\mathrm{S}(\omega) = \mathrm{S}_d(\omega) + \frac{\Delta^2}{12}\frac{2}{f_s}$$

where f_s is the Nyquist sampling frequency. Note how the shaping is limited to a quantisation noise floor of $\frac{\Delta^2}{12}$ as in the classical model of quantisation (CMQ) [9]. In Sec-III, the method in [22] is reconfigured to allow shaping the total error due to quantisation to have a desired $S(\omega)$ via the proper choice of the shape of $S_d(\omega)$ of the dither d.

III. Shaping the PSD of the total error

A block diagram for the proposed method is shown in Fig. 3. A white, unity variance Gaussian process η (with PSD $U(\omega) = 1$) is passed through a properly designed linear time-invariant (LTI) colouring filter $H(j\omega)$ to shape the spectral density $\Phi(\omega) = H(j\omega)H(j\omega)^*U(\omega) = H(j\omega)H(j\omega)^*$ of v. The coloured noise v is then passed through a probability density shaping static non-linearity $\mathcal{F}(\cdot)$ as a composition of the inverse CDF of d and the CDF of v:

$$\mathcal{F}(.) \triangleq F_d^{-1}(F_v(.)). \tag{7}$$

The resulting dither d will have the necessary PDF (in this case a TPDF) and PSD for ε to have a prescribed PSD $S(\omega)$. Since $H(j\omega)$ is a linear filter and $\eta \sim \mathcal{N}(0,1)$, the filtered input will remain Gaussian at the filter output. However, it needs to be scaled by a factor $K \triangleq \frac{\sigma_\eta^2}{\sigma_v^2}$ such that the output also has a unity variance (i.e. $v \sim \mathcal{N}(0,1)$).

The relation between the PSD $\Phi(\omega)$ of v, which can be shaped using $H(j\omega)$, and the desired PSD $S(\omega)$ of ε can be described by means of the corresponding auto-correlation functions (normalised ACFs); $\rho(\tau) \triangleq \frac{\phi(\tau)}{\phi(0)}$ where $\phi(\tau) = E[v(t)v(t+\tau)]$ and $R(\tau) \triangleq \frac{s(\tau)}{s(0)}$ where $s(\tau) = E[\varepsilon(t)\varepsilon(t+\tau)] = E[\varepsilon^2]R(\tau)$. Indeed, from [22] the relation between $R = R(\tau)$ and $\rho = \rho(\tau)$ is as follows:

$$R = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{g(m,n,\rho)}{E[\varepsilon^2]} \\ (Q(\mathcal{F}(m)+x)-x)(Q(\mathcal{F}(n)+x)-x)\,\mathrm{d}m\,\mathrm{d}n. \tag{8}$$

with g the bi-variate Gaussian joint PDF

$$g(m,n,\rho) = \frac{e^{-\frac{1}{2(1-\rho^2)}[m^2+n^2-2\rho mn]}}{2\pi\sqrt{1-\rho^2}}; \tag{9}$$

It is a property of the bi-variate Gaussian joint PDF that $\frac{\partial g}{\partial \rho} = \frac{\partial^2 g}{\partial m \partial n}$. For $Q(\cdot)$ in (4), we therefore get from [22]:

$$\frac{\mathrm{d}R}{\mathrm{d}\rho} = \frac{\Delta^2}{E[\varepsilon^2]}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T}\mathcal{F}'(m)\mathcal{F}'(n)g(m,n,\rho) \\ \chi_{\{\mathcal{F}(m)=T_i-x\}}(m)\chi_{\{\mathcal{F}(n)=T_j-x\}}(n); \tag{10}$$

where $\chi_{\{\}}(\cdot)$ is the indicator function and $\mathcal{F}^{-1}(T_k - x) \triangleq \hat{d}_k$. Note how varying the input value x is captured by a change in the corresponding $\hat{d}_k$. Hence, the solution to (10) is a 3D surface of the variables R, ρ, x; where for each x value, ρ is varied in $(-1,1)$ to find the corresponding R. Now consider $v \sim \mathcal{N}(0,1)$ and d a TPDF dither; this means that $d(t) \in [-\Delta, \Delta]$. Hence $x+d \in [x-\Delta, x+\Delta]$, so for all $x \in (T_2, T_{N_T-1})$ there can only be one or two $\{T_k, T_{k+1}\} \in [x-\Delta, x+\Delta]$ where $p_d \neq 0$). Hence, (10) can be evaluated for the quantisation levels pair $\{T_k, T_{k+1}\}$ when $x \in (T_k, T_{k+1})$ or at the single level T_k when $x = T_k$ for $k \in \{2, 3, ..., N_T - 2\}$.

$$\frac{\mathrm{d}R}{\mathrm{d}\rho} = \sum_{i=k}^{k+1}\sum_{j=k}^{k+1}\frac{4\mathcal{F}'(\hat{d}_i)\mathcal{F}'(\hat{d}_j)}{2\pi\sqrt{1-\rho^2}}e^{-\frac{1}{2}\frac{(\hat{d_i}^2+\hat{d_j}^2-2\rho\hat{d_i}\hat{d_j})}{(1-\rho^2)}} \tag{11}$$

Given the choice of d as a TPDF dither, $\mathcal{F}'(\cdot)$ is:

$$\mathcal{F}'(u) = \Delta\begin{cases}\frac{p_v(\hat{u})}{\sqrt{2F_v(u)}}, & u \in (-\infty, 0] \\ \frac{p_v(\hat{u})}{\sqrt{2(1-F_v(u))}}, & u \in [0, +\infty)\end{cases}. \tag{12}$$

Note how all terms in (11) are non-negative (i.e. $R(\rho)$ is invertable since the solution to (11) is monotone). Taking $R = 0$ when $\rho = 0$, the solution to (11) at $x = T_k$ (i.e. $\hat{d}_k = F_v^{-1}(F_d(0)) = 0$) is:

$$\frac{\mathrm{d}R}{\mathrm{d}\rho} = 4(\mathcal{F}'(0))^2 g(0,0,\rho) = \left(\frac{p_v(0)}{\sqrt{2F_v(0)}}\right)^2\frac{4\Delta^2}{2\pi\sqrt{1-\rho^2}}$$

$$R_{\text{num}} = (\Delta p_v(0))^2\frac{2}{\pi}\sin^{-1}(\rho)$$

$$\text{and so } \rho(\tau) = \sin\left(\frac{\pi}{2}R_\propto(\tau)\right); \tag{13}$$

where

$$R_\propto(\tau) \triangleq \frac{\max(R_{\text{num}}) - \min(R_{\text{num}})}{\max(R) - \min(R)}\frac{R(\tau)}{(\Delta p_v(0))^2}$$

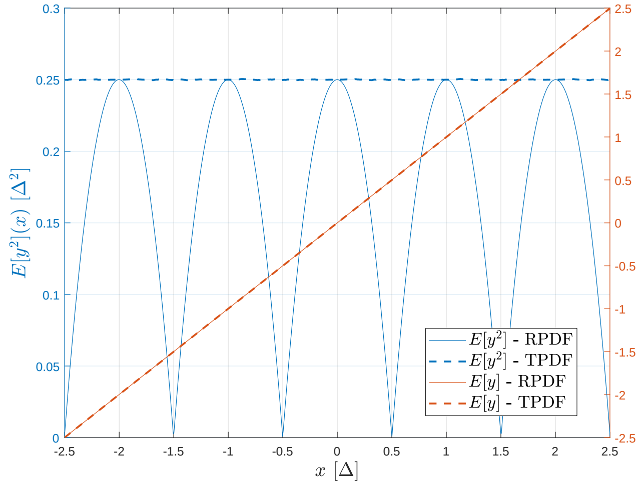

Fig. 4. Dependence of $E[y], E[y^2]$ over x for a 3-Bit uniform quantiser (TPDF vs. RPDF).

is a numerical scaling implemented to assure that both $R(\tau)$ and the corresponding $\rho(R(\tau))$ are in $(-1, 1)$. Note from (13) that R_{num} varies in the feasible numerical range of $(-(\Delta p_v(0))^2, (\Delta p_v(0))^2)$ as ρ varies in $(-1, 1)$. Hence, to obtain $\rho(\tau) \in (-1, 1)$ corresponding to a desired $R(\tau) \in (-1, 1)$, the feasible ranges (of R, R_{num}) are mapped via $R_{\propto}$ (in the case of (13), at $x = T_k$, it simplifies to $R_{\propto}(\tau) = R(\tau)$). Once the mapping $\rho(R)$ can be obtained by (numerically) solving (10) (for a general PDF choice) or (13) (in the case of TPDF) over $\rho \in (-1, 1)$, the effect varying x has over the feasible range of R_{num} values can be investigated. More importantly however, (13) establishes the proper $\rho(\tau)$ corresponding to a desired $R(\tau)$. Note that as long as the specified $s(\tau) = E[\varepsilon^2]R(\tau)$ corresponds to a $\Phi(\omega) \geq 0$ for all ω (PSDs must be non-negative), the desired shaping at the output stage is attainable. Unfortunately, $\Phi(\omega)$ is not guaranteed to be non-negative at all frequencies for every specified $R(\tau)$. This is because when $\rho(\tau)$ in Eq. (8) ranges over all possible normalised ACFs, $R(\tau)$ ranges only over a subset of possible normalised ACFs [22]. So, $H(j\omega)$ is synthesised to best approximate $\Phi(\omega)$ as follows:

$$\text{Let } \Phi^+(\omega) \triangleq \Phi(\omega)\mathcal{X}_{\{\Phi(\omega)\geq 0\}}(\omega), \text{ then}$$

$$H(j\omega) = \sqrt{\Phi^+(\omega)}. \qquad (14)$$

IV. Simulations

In the simulations, the output range is set to $V_{max} = -V_{min} = 10$. Fig. 4 shows the effect of using TPDF over RPDF dither for spectral shaping, as alluded to in the discussion of Sec. II-B. Either dither linearises the quantiser in mean, $E[y] = E[x]$, but TPDF dither also results in a constant total power, subsequently shaped to $S(\omega)$, with $\text{Var}(y) = E[\varepsilon^2] = \Delta^2/4$. The RPDF dither causes a so-called noise modulation effect, i.e. the error power varies with the input x. This effect in the case of RPDF dither can be inferred from Fig. 5, where the range of R_{num} is maximum at $x = T_k = k\Delta$ while it evaluates to zero as it approaches midpoints between quantisation levels, $R_{\propto}(R(\tau)) \to 0$ for all $R(\tau)$.

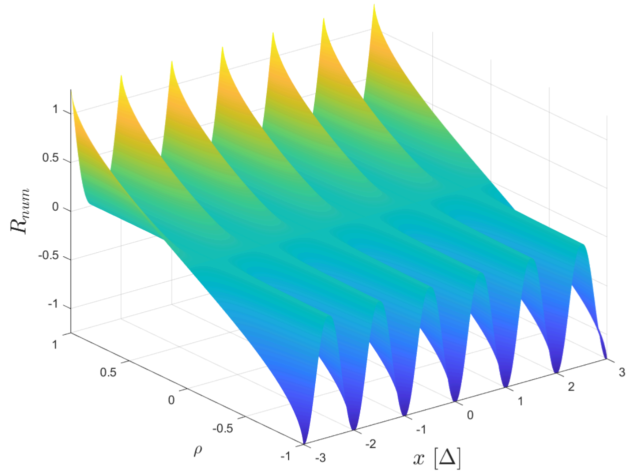

Fig. 5. The dependence of $R(\rho)$ mapping on the input x in the case of RPDF dither. Solving (10)-(13) for $R_{\text{num}}(\rho, x) : \rho \in (-1, 1), x \in [V_{min} + \frac{\Delta}{2}, V_{max} - \frac{\Delta}{2}]$ for a 3-Bit uniform quantiser.

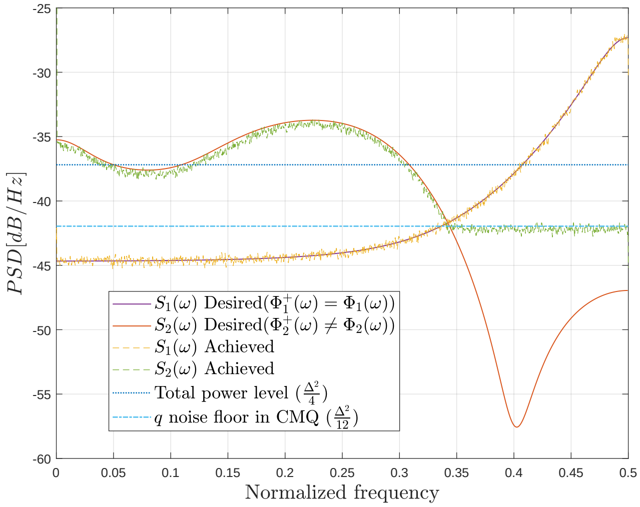

Fig. 6. Desired vs. Achieved single-sided output PSD $S(\omega)$ for TPDF NSD of a 10−Bit uniform quantiser.

Algorithm 1 was used to generate randomised PSDs with arbitrary shapes. Desired error PSDs must be feasible, i.e. $S(\omega) = S(-\omega) \in \mathbb{R}_{\geq 0}$ and scaled to have an area equivalent to an achievable total power at the output, $\int_{\mathbb{R}} S(\omega)\,\mathrm{d}\omega = E[\varepsilon^2] = \Delta^2/4$. Figs. 6 and 7 show results of the method when shaping the PSD of ε to a desired $S(\omega)$.

V. Results and Discussion

In order to get an indication of the range of PSDs that can be realised using the method, 3000 randomised PSD function samples were generated as described in Sec. IV. 68% of the generated PSDs were not entirely non-negative, $\Phi^+(\omega) \neq \Phi(\omega)$, indicating that there are limitations on the properties that a PSD function can exhibit in order to be fully

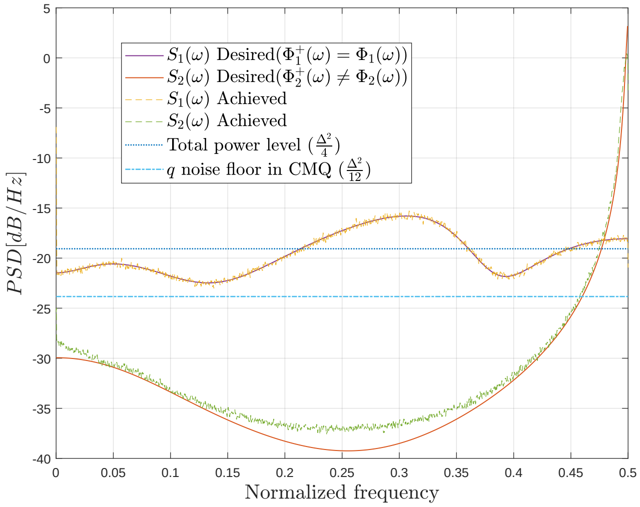

Fig. 7. Desired vs. Achieved single-sided output PSD $S(\omega)$ for TPDF NSD of a $7-$Bit uniform quantiser.

Algorithm 1 PSD Generator, MATLAB

```
G = drss(7); % Random discrete State Space, Stable SISO
 G_tf = tf(G);
 S_tf = G_tf*G_tf'; % Real valued, non-negative, PSD
 S_num = S_tf.Numerator{:};
 S_den = S_tf.Denominator{:};
w = linspace(0,2*pi,M); % Sample whole circle
 S_fr = freqz(S_num,S_den,w); % Frequency response at w
 S_fr_2norm = sum(abs(S_fr).*mean(diff(w)))/(2*pi);% 2-Norm
% Normailze area S(w) to a unity power
 S_fr_ = (S_fr/S_fr_2norm)*Var_y; % Scale to feasible power
```

realised using the proposed method. Of the 68%, 72% of the PSDs were not non-negative for ω where $S(\omega) < \Delta^2/12$, exemplified by $S_2(\omega)$ in Fig. 6.

Of the 32% PSDs resulting in $\Phi^+(\omega) = \Phi(\omega)$, 94% produced the desired $S(\omega)$. Of these, 75% of the cases satisfied $S(\omega) > \Delta^2/12$ for all ω, exemplified by $S_1(\omega)$ in Fig. 7. The remaining 6% had a deviation from $S(\omega)$ when $S(\omega)$ had a concentration of power in a narrow band of frequencies, or there was a mismatch when $S(\omega)$ dropped approximately 8-dB below the CMQ noise floor, in the 8-bit quantiser case. For higher number of bits the spectral density tended to reach the CMQ noise floor. In general, the results indicate that PSDs that drop below the CMQ noise floor often have a mismatch between desired and achieved results, but the CMQ noise floor apparently is not a general lower bound, as exemplified by $S_1(\omega)$ in Fig. 6 and $S_2(\omega)$ in Fig. 7. These examples indicate that increased performance in terms of spectral shaping of the total error via NSD is possible, as compared to the results in Sec.II-C & Sec.III-E in [8]. Future work will focus on determining the limitations and bounds of the proposed spectral shaping method.

VI. Conclusions

It was demonstrated by way of simulation that the proposed dithering method can provide a large degree of freedom in shaping the spectral distribution of the total error power for a uniform quantiser. The choice of a triangular probability density for the dither makes the error power independent of the input, enabling shaping of the error power spectral density. This requires a jointly specified probability and spectral density for the dither, and the required results were developed in order to generate such a dither. Simulation results indicate that there are limitations to achievable spectral densities, and these should be investigated further. However, the method can in its current form produce improved performance compared to existing non-subtractive dithering methods.

References

[1] R. E. Crochiere and L. Rabiner, "Interpolation and decimation of digital signals—A tutorial review," in *Proc. IEEE*. IEEE, 1981, pp. 300–331.

[2] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical Theory of Quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, 1996.

[3] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. IEEE Press, 2005.

[4] M. Neitola, "Lee's Rule Extended," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 4, pp. 382–386, 2017.

[5] G. C. Goodwin, D. E. Quevedo, and D. McGrath, "Moving-horizon optimal quantizer for audio signals," *Journal of the Audio Engineering Society*, vol. 51, no. 3, pp. 138–149, 2003.

[6] B. Adhikari, R. van der Rots, J. Leth, and A. Eielsen, "Linearisation of digital-to-analog converters by model predictive control," in *IFAC Conference on Nonlinear Model Predictive Control*, 2024.

[7] N. van Rijt, A. Faza, T. Oomen, and A. A. Eielsen, "Learning control applied to a digital-to-analogue converter," in *IEEE Conference on Control Technology and Applications*. IEEE, 2023, pp. 91–96.

[8] R. A. Wannamaker, S. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Nonsubtractive Dither," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 499–516, 2000.

[9] B. Widrow and I. Kollár, *Quantization Noise*. Cambridge University Press, 2008.

[10] A. A. Eielsen and A. J. Fleming, "Improving Digital-to-Analog Converter Linearity by Large High-Frequency Dithering," *IEEE Trans. Circuits Syst. I*, vol. 64, no. 6, pp. 1409–1420, 2017.

[11] C. A. Desoer and S. M. Shahruz, "Stability of dithered non-linear systems with backlash or hysteresis," *International Journal of Control*, vol. 43, no. 4, pp. 1045–1060, 1986.

[12] B. Armstrong-Hélouvry, P. Dupont, and C. C. De Wit, "A survey of models, analysis tools and compensation methods for the control of machines with friction," *Automatica*, vol. 30, no. 7, pp. 1083–1138, 1994.

[13] A. A. Eielsen, J. Leth, A. J. Fleming, A. G. Wills, and B. Ninness, "Large-amplitude dithering mitigates glitches in digital-to-analogue converters," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1950–1963, 2020.

[14] A. Faza, J. Leth, and A. A. Eielsen, "Mitigating non-linear dac glitches using dither in closed-loop nano-positioning applications," in *2023 American Control Conference (ACC)*. IEEE, 2023, pp. 685–691.

[15] ——, "Criterion for sufficiently large dither amplitude to mitigate non-linear glitches," in *IEEE Conference on Control Technology and Applications*. IEEE, 2023, pp. 970–977.

[16] B. Widrow and I. Kollar, *Quantization noise roundoff error in digital computation, signal processing, control, and communications*. Cambridge University Press, 2008.

[17] D. E. Knuth, *The Art of Computer Programming*, 3rd ed. Addison-Wesley, 1997, vol. 2.

[18] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*. Wiley-Interscience, 1997.

[19] A. Papoulis, *The Fourier Integral and Its Applications*. McGraw-Hill, 1962.

[20] G. Wise, A. Traganitis, and J. Thomas, "The effect of a memoryless nonlinearity on the spectrum of a random process," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 84–89, 1977.

[21] U. Gujar and R. Kavanagh, "Generation of random signals with specified probability density functions and power density spectra," *IEEE Transactions on Automatic Control*, vol. 13, no. 6, pp. 716–719, 1968.

[22] M. M. Sondhi, "Random processes with specified spectral density and first-order probability density," *Bell System Technical Journal*, vol. 62, no. 3, pp. 679–701, 1983.