Filtering of subtractive discrete dither in quantizers: some new results

Abhishek Ghosh, *Student Member, IEEE* and Sudhakar Pamarti, *Member, IEEE*
Department of Electrical Engineering, University of California, Los Angeles, CA 90095
Email:{abhishek,spamarti}@ee.ucla.edu

***Abstract*—Subtractive dither in quantizers is examined as a means to mitigate quantizer non-linearity. The effects of filtering the dither signal to shape its spectral content outside the signal band while maintaining its benefits are studied in detail. Design strategies for finite impulse response (FIR) filters that accomplish spectral shaping as well as allay quantizer non-linearity are derived theoretically. Simulation results for low/medium resolution quantizers are presented to validate the derived conditions on the filter structures.**

I. Introduction

Quantizers are the portals to digital signal processing of all real-world signals and hence serve as the main interface between natural and machine-based signal processing. An example mid-tread quantizer is shown in Fig. 1. As can be seen, the input-output characteristic of any example quantizer, is non-linear and hence signals when *quantized* result in errors which have significant dependence on the input signal and hence are spectrally non-white [1]–[3], [5]. A major understanding from these works is that the input signal to the quantizer needs to be equipped with certain statistical properties in order to ensure that the quantization error is white and its power is the ubiquitous $\Delta^2/12$ (Δ being the quantization step size). In most practical scenarios though, it is highly infeasible to handle signals with the required statistical properties. So, a small signal, random in nature (called *dither*) is added to the input in order to make the composite signal samples *unpredictable* at any given time.

A. Dithered quantization

Let us define a dithered quantizer more formally. A behavioral schematic is presented in Fig. 2. A random signal $r[n]$ is added to the signal to be quantized $x[n]$ and the composite signal $z[n] = x[n] + r[n]$ is passed through the quantizer. A dithered quantizer can be implemented in a few different flavors, each unique in the properties it imparts to the quantization error, Figs. 2(a)-(c).

In Fig. 2(a), the added dither signal $r[n]$ is subtracted digitally from the quantized value $y[n]$ and hence is called a *subtractively dithered quantizer*. Likewise, Fig. 2(b) refers to a non-subtractively dithered quantizer (commonly phrased as additive dithered quantizer). The added dither, $r[n]$ is usually constrained to be bounded between one least significant bit (LSB) of the quantizer. Separate conditions [3] have been theoretically derived for either case to ensure that the error-samples ($e[n] = y[n]-z[n]$ for Fig. 2(a) and $e[n] = y[n]-x[n]$ for Fig. 2(b)) are independent (among themselves as well as of the input) and uniformly distributed both in terms of first and second order statistics. Henceforth, such error would be called *well-behaved* in this paper. It is found that a well-behaved quantization error can be guaranteed, for a subtractive dithered quantizer, if the dither statistics satisfies certain properties. The simplest class of dither conforming to these statistics is a uniformly distributed dither. However, for an additive dither situation, such conditions are not easily derivable [3].

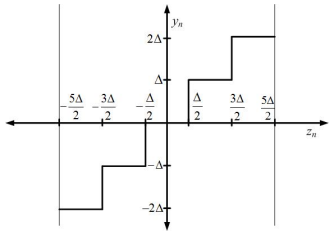


Fig. 1: Mid-tread quantizer

Unfortunately, a uniformly distributed white dither signal, $r[n]$ would contribute too much noise to the quantizer output. In fact, a uniformly distributed dither signal spanning one quantizer LSB would degrade the overall signal-to-noise ratio (SNR) by 3dB. Furthermore, it may be impossible or at least extremely challenging to digitally generate such analog dither [6]. The second problem is solved using hardware-friendly digital dither (that spans only a finite set of values), while the first problem is solved by spectrally shaping such dither out of the band of interest using digital filters [4]. Such an architecture is presented in Fig. 2(c) as an extension of Fig. 2(a), where $d[n]$ is a Bernoulli signal with equal probability of a 0 or 1. However, filtering a signal tantamounts to modifying its statistical properties. Consequently, the error signal $e[n]$ in Fig. 2(c) may not be well-behaved as noted above even if $d[n]$ is sample-wise independent, identically distributed (i.i.d.) and white.

B. Prior-art

There have been some very interesting works treating filtered dither signals and their efficacies in whitening the quantization error [4], [7], [8]. With reference to Fig. 2(c), in [4], a detailed analysis is done on the properties of $r[n]$ where $d[n]$'s are i.i.d. random variables. However, the analysis is specific to additive dithered quantizers and imposes very strict

conditions on the filter-coefficients (finite (FIR) or infinite (IIR) impulse response). In [8], a simplified condition is derived for FIR filters. However, the quantizer treated in [8], works on integer valued inputs only while the work in [7] provides conditions for the impulse response of an IIR filter (integrator in feed-forward path of a sigma-delta modulator).

C. Contribution

The main contribution of this work is the formulation of conditions for integer-valued FIR filters operating on continuous-valued inputs for a subtractively dithered quantizer. We theoretically derive conditions for the error-sequence, subject to such filtering, to be well-behaved. In the next section, we detail the behavioral model to be used in all subsequent derivations. Section III furnishes the main theoretical results accompanied by some relevant proofs while Section IV provides an insight into the simulation results to validate the theory. We conclude the paper in Section V.

II. Behavioral model

The model is as presented in Fig. 2(c). Let us define an i.i.d. Bernoulli sequence $d[n]$ that follows the statistics: $\mathbb{P}r(d[n] = 0) = \mathbb{P}r(d[n] = 1) = 0.5$. The sequence $d[n]$ is passed through a digital filter $G(z)$ having a finite impulse response $g[n] \in \mathbb{Z}$ of length K to produce an output $r[n]$. The filter gain is so adjusted that the output $r[n]$ spans Δ, LSB of the quantizer. Consequently, the filtered output $r[n]$ can be expressed as

$$\begin{aligned} r[n] =& \frac{\Delta}{L}(g[0]d[n] + g[1]d[n-1] + ... \\ & .... + g[K-1]d[n-K+1]) \end{aligned} \tag{1}$$

where $L = \sum_{i=0}^{K-1} |g[i]|$. The quantity Δ/L can be thought of as the dither LSB (the minimum resolution of the added signal r_n). The input signal $x[n]$ is assumed to be of arbitrary distribution and bounded in $[-(Q-1)\Delta/2, (Q-1)\Delta/2]$ for a Q-level quantizer ($Q \in \mathbb{N} \cap (1, \infty)$). The signal $r[n]$ is added with the input $x[n]$ to result in the composite signal $z[n] = x[n] + r[n]$. $z[n]$ is quantized to generate $v[n]$. The added dither signal $r[n]$ is subtracted from $v[n]$ to result in the actual output $y[n]$. The resultant quantization error is defined as $e[n] = y[n] - x[n] = v[n] - x[n] - r[n]$.

Note-1: Since the dither resolution is finite, namely Δ/L, hence any input of the form $x[n] = [x[n]] + \langle x[n] \rangle$ where $[x[n]] = k\Delta, k \in \mathbb{Z}$ and $\langle x[n] \rangle < \Delta/L$ will not *see* the effect of the added dither, and hence the quantization error will not be guaranteed to be well-behaved. In the remainder of the paper, we shall assume that the input signal $x[n]$ excludes the above special class of signals.

Note-2: In the following arguments, w_j and $w[j]$ would refer to the same quantity and will be used interchangeably.

III. Main result: theory

Theorem 1: For a dithered quantizer, modeled in Section II,

P.1) The error sequence e_n is an identically distributed uniform random variable independent of the input $x_{n-m}, \forall k_1 \in \mathbb{Z}, \forall m \in \mathbb{Z}$ if and only if ($\langle\rangle_T$ operator denotes modulo-T operation)

C.1) A non-negative integer $i < K$ exists such that $\langle g_i k_1 \rangle_L = L/2$

P.2) The error sequence pair $(e_n, e_{n-p}) \forall p \in \mathbb{Z} \cap (0, K)$ is pairwise independent, each being an identical uniform distribution $\forall (k_1, k_2) \neq (0, 0)$ if and only if either of the following are true

C.2) A non-negative integer $l < p$ exists such that $\langle g_l k_1 \rangle_L = L/2$

C.3) A non-negative integer $1 \leq r \leq p$ exists such that $\langle g_{K-r} k_2 \rangle_L = L/2$

C.4) A non-negative integer $p \leq m < K$ exists such that $\langle g_m k_1 + g_{m-p} k_2 \rangle_L = L/2$

P.3) The error sequence pair $(e_n, e_{n-p}) \forall p \in \mathbb{Z} \cap [K, \infty)$ is pairwise independent, each being an identical uniform distribution $\forall (k_1, k_2) \neq (0, 0)$ if both the following conditions hold

C.5) The FIR filter coefficients $g[k]$ are of the form 2^i where i takes on each value in $[0, s-1]$ at least once and

C.6) $L = \sum_{i=0}^{K-1} |g[i]| = 2^s$ where $s \in \mathbb{Z} \cap (1, K]$

Remark: For notational convenience, all properties are denoted as P.'s while all conditions are denoted as C.'s. Both P.1 and P.2 are *if and only if* conditions while P.3 is only a sufficiency condition. The strategy of the proof would be to proceed with P.2 first. The proof of P.1 would follow next while P.3 would be proved as a consequence of P.2 and would form the main result of this work, providing easy-to-use closed form solutions for the shaping filter $G(z)$. Let us proceed with P.2 now.

Proof: The proof would use characteristic functions [9] to derive conditions on the specific properties of the added dither signal. This is a commonly used technique for such applications [8]. In fact, from [3], we know, that the joint characteristic function (jcf) for error-samples (e_n, e_{n-p}) can be written as, $\forall p \in \mathbb{Z} \cap (0, K)$ for $(k_1, k_2) \in \mathbb{Z}^2$

$$\begin{aligned} \Phi_{e_n, e_{n-p}}(u_1, u_2) = & \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \frac{\sin(\pi\Delta(u_1 - k_1/\Delta))}{(\pi\Delta(u_1 - k_1/\Delta))} \\ & \frac{\sin(\pi\Delta(u_2 - k_2/\Delta))}{(\pi\Delta(u_2 - k_2/\Delta))} \\ & \Phi_{x_n, x_{n-p}}(\frac{-2\pi k_1}{\Delta}, \frac{-2\pi k_2}{\Delta}) \\ & \Phi_{r_n, r_{n-p}}(\frac{-2\pi k_1}{\Delta}, \frac{-2\pi k_2}{\Delta}) \end{aligned} \tag{2}$$

Hence, for the joint density of (e_n, e_{n-p}) to be uniform and pairwise independent, it suffices to show [3],

$$\begin{aligned} \Phi_{r_n, r_{n-p}}(\frac{-2\pi k_1}{\Delta}, \frac{-2\pi k_2}{\Delta}) = 0 \\ \forall (k_1, k_2) \in \mathbb{Z}^2 - (0, 0) \end{aligned} \tag{3}$$

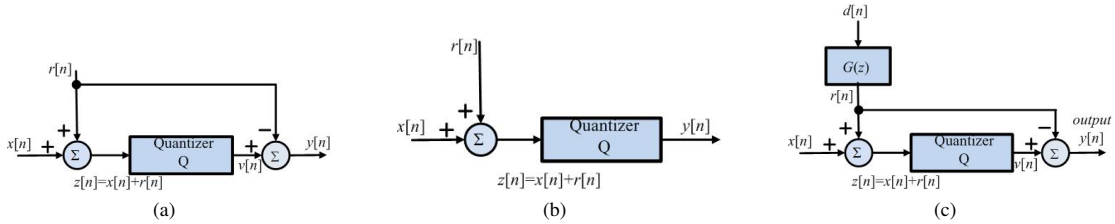


Fig. 2: Dithered Quantizers: (a) Subtractive (b) Non-subtractive/Additive (c) Filtered-subtractive

The jcf of the dither samples (r_n, r_{n-p}) is defined as

$$\begin{aligned}\Phi_{r_n,r_{n-p}}(u_1,u_2) &= \mathbb{E}(e^{j(u_1 r_n + u_2 r_{n-p})} \\ &= \mathbb{E}(e^{j\frac{\Delta}{L}(u_1\sum_{m=0}^{K-1} g_m d_{n-m} + u_2 \sum_{l=0}^{K-1} g_l d_{n-p-l})} \\ &= \prod_{l=0}^{p-1} \Phi_d(\frac{\Delta}{L}u_1 g_l) \\ &\prod_{m=p}^{K-1} \Phi_d(\frac{\Delta}{L}(u_1 g_m + u_2 g_{m-p})) \\ &\prod_{r=1}^{p} \Phi_d(\frac{\Delta}{L}u_2 g_{K-r}) \end{aligned} \tag{4}$$

Now, for a Bernoulli dither d_n, with $\mathbb{P}r(d_n = 0) = \mathbb{P}r(d_n = 1) = 0.5$,

$$\Phi_d(v) = e^{(-jv/2)}\cos(v/2) \tag{5}$$

From Eqn. (3), we need to evaluate Eqn. (4) for $u_{1,2} = 2\pi k_{1,2}/\Delta$. Thus, from Eqns. (4) and (5), we can write $\forall (k_1,k_2) \in \mathbb{Z}^2 - (0,0)$

$$\begin{aligned}|\Phi_{r_n,r_{n-p}}(-\frac{2\pi k_1}{\Delta}, \frac{-2\pi k_2}{\Delta})| &= \prod_{l=0}^{p-1} |\cos(\frac{\pi k_1 g_l}{L})| \\ &\prod_{m=p}^{K-1} |\cos(\frac{\pi(k_1 g_m + k_2 g_{m-p})}{L})| \\ &\prod_{r=1}^{p} |\cos(\frac{\pi k_2 g_{K-r}}{L})| \end{aligned} \tag{6}$$

This proves the *sufficiency* of the theorem, since if any one of the product series terms is zero (C.2-4), P.2 is satisfied.

Necessity: The necessity conditions can be likewise argued, and is omitted here for brevity.

Discussion: It may not be always possible to design FIR filter coefficients satisfying conditions C.2-4 of Theorem 1 since the filter coefficients are not available in a closed-form solution. Furthermore, it's not practically possible to evaluate the characteristic function in Eqn (6). at all integer values of (k_1,k_2) to identify an appropriate filter structure. P.3 addresses this issue in further detail.

For the proof of P.1, we write the probability density function (pdf) of the error sequence e_n conditioned on the input $x_{n-m} \forall m \in \mathbb{Z}$ as

$$p_{e_n|x_{n-m}}(a|b) = \sum_{l=-\infty}^{\infty} p_{z_n|x_{n-m}}(-a + l\Delta|b) \tag{7}$$

Now, it is not difficult to see that

$$\begin{aligned} p_{z_n|x_{n-m}}(c|b) &= p_{x_n+r_n|x_{n-m}}(c|b) \\ &= p_{x_n|x_{n-m}}(c|b) * p_{r_n}(c) \end{aligned} \tag{8}$$

since r_n is independent of both x_n and x_{n-m} where a, b and c are in the appropriate domains and $*$ denotes convolution.

Thus, from Eqns. (7) and (8), the characteristic function (cf) of e_n conditioned on x_{n-m}, can be written as

$$\begin{aligned}\Phi_{e_n|x_{n-m}}(u) = \frac{1}{\Delta}\sum_{k=-\infty}^{\infty} &\Phi_{x_n|x_{n-m}}(-u) \\ &\prod_{i=0}^{K-1} \Phi_{d_n}(-u g_i \frac{\Delta}{L}) \frac{\sin(\pi\Delta(u_1 - k_1/\Delta))}{(\pi\Delta(u_1 - k_1/\Delta))} \end{aligned} \tag{9}$$

For the error-sequence e_n to be independent of x_{n-m} and be uniformly identically distributed, $\Phi_{e_n|x_{n-m}}(-\frac{2\pi k_1}{\Delta})$ must evaluate to 0 for every $k_1 \neq 0$. For this to happen, for any arbitrary input, from the proof of P.2,

$$\begin{aligned}\prod_{i=0}^{K-1} |\Phi_{d_n}(\frac{-2\pi k_1 g_i \frac{\Delta}{L}}{\Delta})| &= \prod_{i=0}^{K-1} |\Phi_{d_n}(\frac{-2\pi k_1 g_i}{L})| \\ &= \prod_{i=0}^{K-1} |\cos(\frac{\pi k_1 g_i}{L})| = 0 \end{aligned} \tag{10}$$

Eqn. 10 holds if and only if C.1 holds (the argument of at least one cosine term is driven to an odd multiple of $\pi/2$) hence proving P.1

The proof of P.3 will lead from that of P.2 through an important observation. Since, $p \geq K$, hence it is not difficult to see that,

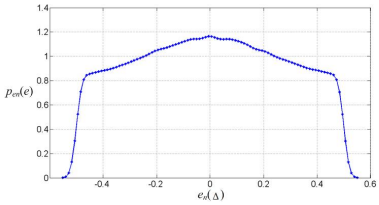
$$\begin{aligned} p_{r_n,r_{n-p}}(r_1,r_2) &= p_{r_n}(r_1)p_{r_{n-p}}(r_2) \\ \Phi_{r_n,r_{n-p}}(u_1,u_2) &= \Phi_{r_n}(u_1)\Phi_{r_{n-p}}(u_2) \end{aligned} \tag{11}$$

Now, from Eqn. (3), we need to prove that $\Phi_{r_n,r_{n-p}}(\frac{-2\pi k_1}{\Delta}, \frac{-2\pi k_2}{\Delta})$ goes to zero for all values of $(k_1,k_2) \in \mathbb{Z}^2 - (0,0), \forall p \in \mathbb{Z} \cap [K,\infty)$. Based on Eqns. (4)-(6), this is equivalent to proving

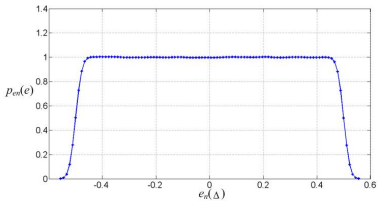
$$\prod_{i=0}^{K-1} |\cos(\frac{\pi k_1 g_i}{L})||\cos(\frac{\pi k_2 g_{K-1-i}}{L})| = 0 \tag{12}$$

for all values of $(k_1,k_2) \in \mathbb{Z}^2 - (0,0), \forall p \in \mathbb{Z} \cap [K,\infty)$.

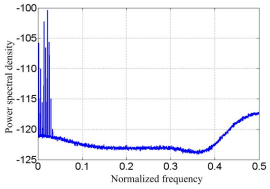
It is interesting to note that Eqn. (12) leads to an L-periodic sequence (in k_1 or k_2) if condition C.1 is satisfied. Consequently, it suffices to evaluate the cf of Eqn. (12) in a finite set of L^2 points. Now it becomes useful to consider the



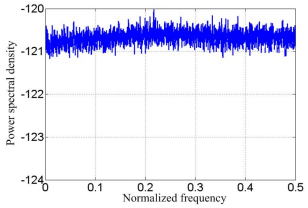
(a) error pdf for G_1



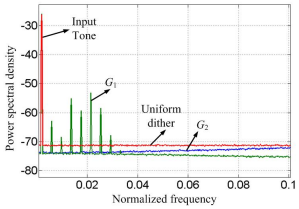
(b) error pdf for G_2



(c) error psd for G_1



(d) error psd for G_2



(e) Output spectrum for three different scenarios

Fig. 3: Simulation results

following cases, $\forall(k_1, k_2) \in [-L/2+1, L/2]$ assuming the conditions in Theorem 1 hold (sufficiency).

- $k_1 = odd, k_2 = odd$ One product term of the right-hand side of Eqn. (12) can be written as $\cos(\pi k_j \frac{2^r}{2^s}), j = 1, 2$. Hence for $r = s-1$, we can write the product term as $\cos(\frac{\pi}{2}k_j)$ which goes to 0 since $k_{1,2}$ are odd.
- $k_1 = odd, k_2 = even$ Here, k_1 will drive the product term to 0 for $r = s-1$. The symmetric case of $k_2 = odd, k_1 = even$ similarly can be shown to equate to 0.
- $k_1 = even, k_2 = even$ Here, let $k_{1,2} = 2^l(2m+1), l \leq s-1$ for any integer m. Then the product term containing $r = s-1-l$ would yield $\cos(\frac{\pi}{2}(2m+1))$ which again goes to 0.

■

Discussion: C.5 and C.6 give easy formulae to design the dither-shaping filter. The proposed solution to Eqn. (12) may not be unique (under investigation), but the aforementioned conditions are in tune with powers-of-2 FIR filters [10], and hence amenable to facile design. It should further be noted that condition C.5 is a subset of C.1 and hence ensures an uniformly distributed error sequence independent of the input. It is of interest to observe that though P.5 proves pair-wise independence only for error samples separated by more than K, for all practical purposes the error is white with a uniform distribution.

IV. Main results: simulation

Let us consider two filters, $G_1(z)$ and $G_2(z)$ (z-transforms of 2 example filters $g_1[n]$ and $g_2[n]$ respectively) such that the former satisfies neither of C.5 and C.6 while the latter satisfies both.

$$G_1(z) = 1 - 3z^{-1} + 5z^{-2} - 9z^{-3} + 3z^{-4} - 3z^{-5} + 9z^{-6} - 5z^{-7} + 3z^{-8} - z^{-9}$$
$$G_2(z) = -1 - 2z^{-1} - 4z^{-2} - 8z^{-3} + 16z^{-4} - z^{-5}$$

The input $x[n]$ is chosen to be a continuous-valued sinusoid at a normalized frequency of 0.002 with an amplitude of 2Δ. The signal is quantized into $Q = 5$ levels as in Fig. 1. In Fig. 3(a),(b), we plot the pdf of the error sequence e_n for both the cases, while Fig. 3(c)(d) shows the spectra of the error signal. As can be clearly seen, the proposed filter, namely G_2, whitens the error-sequence and exhibits an almost uniform pdf (Fig. 3(b)) while G_1 shows an almost triangular pdf (Fig. 3(a)) for the error samples. The error power spectral density (psd) for G_2 (Fig. 3(d)) is *white*, while the error psd for G_1 exhibits multiple spurious tones at harmonic frequencies (as is expected from a lookup table type non-linearity) (Fig. 3(c)). In order to make a fair comparison, a third case where a uniform dither signal $r[n]$ (the case in Fig. 2(b)) is added to the input signal before quantizing, is also considered. The spectra of $y[n] = x[n] + e[n]$ is plotted for all the three cases: G_1, G_2 and uniform dither in Fig. 3(e). As can be seen, the uniform dithered quantizer contributes the maximal in-band power while whitening the output spectrum completely. G_2 shapes the in-band dither power, as well as gets rid of any spurious components, while G_1 has the least in-band dither power contribution but engenders harmful spurious tones at the quantizer output. This is expected since, from P.1 $e[n]$ being independent of $x[n]$ bequeaths the well-behaved properties of $e[n]$ on $y[n]$.

V. Conclusion

A filtered dithering technique in quantizers is proposed. Theoretical conditions on the filter structure are derived to ensure independence, whiteness and uniform distribution of the quantization error signal. Behavioral simulation results are presented to corroborate the proposed results and claims.

References

[1] Sripad, A.; Snyder, D.; , "A necessary and sufficient condition for quantization errors to be uniform and white," *Acoustics, Speech and Signal Processing, IEEE Transactions on* , vol.25, no.5, pp. 442- 448, Oct 1977

[2] Widrow, B.; Kollar, I.; , "Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications," *Cambridge University Press*, Cambridge, UK

[3] Lipshitz, S.P.; Wannamaker, R.A.; Vanderkooy, J.; Wright, J.N.; , "Quantization and dither: a Theoretical Survey," in *J.AudioEng.Soc.*, ,Vol.40, no.5, May 1992.

[4] Wannamaker, R.A.; Lipshitz, S.P.; Vanderkooy, J.; , "Dithered quantizers with and without feedback," *Applications of Signal Processing to Audio and Acoustics, 1993. Final Program and Paper Summaries., 1993* IEEE Workshop on , vol., no., pp.140-143, 17-20 Oct 1993

[5] Gray, R.M.; Stockham, T.G., Jr.; , "Dithered quantizers," *Information Theory, IEEE Transactions on* , vol.39, no.3, pp.805-812, May 1993

[6] Borkowski, M.J.; Kostamovaara, J.; , "On randomization of digital delta-sigma modulators with DC inputs," *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on* , vol., no., pp.4 pp., 21-24 May 2006

[7] Pamarti, S.; Welz, J.; Galton, I.; , "Statistics of the Quantization Noise in 1-Bit Dithered Single-Quantizer Digital DeltaŰSigma Modulators," *Circuits and Systems I: Regular Papers, IEEE Transactions on* , vol.54, no.3, pp.492-503, March 2007

[8] Pamarti, S.; Delshadpour, S.; , "A Spur Elimination Technique for Phase Interpolation-Based Fractional- N PLLs," *Circuits and Systems I: Regular Papers, IEEE Transactions on* , vol.55, no.6, pp.1639-1647, July 2008

[9] Papoulis, A.; Pillai, S.; , "Probability, Random Variables and Stochastic Processes," *McGraw Hill*, 2001

[10] Y. Lim and S. Parker, "FIR filter design over a discrete powers-of-two coefficient space," *IEEE Trans. Acoust., Speech, Signal Process*, vol.31, no.3, pp.583Ű591, 1983