14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

Roundoff Noise Analysis of Signals Represented Using Signed Power-of-Two Terms

Ya Jun Yu *and* Yong Ching Lim
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798

***Abstract*— Multiplierless filtering is very attractive for digital signal processing, for the coefficient multiplier is the most complex and the slowest component. An alternative way to achieve multiplierless filtering other than designing digital filters with signed power-of-two (SPT) coefficient values is to round each input data to a sum of a limited number of SPT terms. However, a roundoff noise representing the roundoff error is introduced when signal data are rounded to SPT numbers. In the SPT space, the quantization step size is nonuniform and thus the roundoff noise characteristic is different from that produced when the quantization step size is uniform. This paper presents an analysis for the roundoff noise of signal represented using a limited number of SPT terms. Roundoff errors for Gaussian distributed inputs are estimated by using our analysis. Examples show that the estimated errors are very close to the actual ones.**

I. Introduction

It is a well known fact that the multiplication of a number by an integer power-of-two is a very simple process in binary arithmetic. Hence, digital filters whose coefficient values are integer power-of-two are essentially multiplierless. However, the design of digital filters with signed power-of-two (SPT) coefficient values require time consuming optimization process and may not always be possible in some applications such as in adaptive filtering. Since hardware circuitry for real time conversion of a binary integer into a limited number of SPT is available [1], if the signal is expressed in SPT terms, i.e., in digit code, the filter is also multiplierless even though the coefficient values are not SPT.

When each signal data is rounded to a limited number of SPT terms, a roundoff noise representing the roundoff error is introduced. In the SPT space, the quantization step size is nonuniform [2] and thus the roundoff noise characteristic is different from that produced when the quantization step size is uniform. Whereas the finite wordlength roundoff error is most frequently statistically modeled as an additive, uniformly distributed white noise [3], the understanding on the statistics of SPT roundoff error, so far, is very limited. Some statistical analysis on the SPT number had been reported in [2], [4]. Lim et al [2] investigated the relationships between SPT terms and the number they represent. These relationships lead to a statistical measure on the number of SPT terms required to represent an integer. Yao [4] derived the distribution of the SPT terms, i.e., the probability of allocating an SPT term to a particular digit. Bussgang [5] investigated the autocorrelation and crosscorrelation functions of Gaussian signal after they undergo nonlinear amplitude distortion. However, there is no report on the statistical distribution of SPT quantization error for the rounding procedure.

The purpose of this paper is to present an analysis for the roundoff noise of signal represented using a limited number SPT terms. Section II reviews some properties of SPT numbers; these properties will be used in the subsequent derivations. In Section III, an SPT roundoff error model is established for a given distributed input. Mathematical expressions on the error's probability density function are established subject to a given precision, and the statistical properties of the roundoff error are studied. Based on this SPT roundoff error model, in Section IV, the roundoff error for Gaussian distributed inputs is approximated.

II. Signed Power-of-Two Numbers

A number n can be represented to a precision 2^Q by an $(L-Q)$-digit canonic SPT number with K SPT terms as

$$n=\sum_{i=0}^{K-1} y(i)2^{q(i)}, y(i)\in\{-1,1\}, \tag{1}$$

where $Q\le q(i)\le L-1$. $L-Q$ is the wordlength of the SPT number being a positive integer but Q may be zero, a positive integer, or a negative integer. Furthermore, for any i and j, it satisfies the constraints that $q(i)\neq q(j)$ if $i\neq j$ and that $q(i)\neq q(j)+1$.

For the particular condition in (1) where $Q=0$, n is an L-digit integer. In this section, this particular case for quantizing a number to an SPT number where $Q=0$ is considered.

Let $\mathcal{Z}$ denote the set of all integers and $\mathcal{Z}^+$ denote the set of all positive integers. Let $L,K\in\mathcal{Z}^+$. Since $q(i)\neq q(j)+1$, it is obvious that an L-digit canonic SPT integer has at most $\left\lfloor\frac{L+1}{2}\right\rfloor$ SPT terms (or non-zero digits), where $\lfloor x\rfloor$ is the largest integer smaller than or equal to x. Assume further that $L\ge 2K-1$. Let $T^+(L,K)$ be a subset of $\mathcal{Z}^+$ such that any $n\in T^+(L,K)$ is a sum of exactly K canonic SPT terms and the largest power-of-two term is less than or equal to 2^{L-1}, where n is given by (1) in which Q is equal to 0.

It is known that the number of elements of the set $T^+(L,K)$, represented as $N^+(L,K)$, is [2]

$$N^+(L,K)=\frac{2^{K-1}}{K!}\prod_{k=0}^{K-1}(L-K+1-k). \tag{2}$$

In (2), $K!$ denotes the factorial of K.

Let $S^+(L,K)$ be a subset of $\mathcal{Z}^+$ such that $S^+(L,K)=\bigcup_{k=1}^{K}T^+(L,k)$, i.e., any $n\in S^+(L,K)$ is a sum of not more than K canonic SPT terms and the largest power-of-two term is less than or equal to 2^{L-1}. It is noted that $S^+(L,K)$ does not include zero.

Let $M^+(L,K)$ be the number of elements of the set $S^+(L,K)$. It is straightforward to show that

$$M^+(L,K)=\sum_{k=1}^{K}N^+(L,k). \tag{3}$$

Therefore, the number of elements of $S^+\left(L,\left\lfloor\frac{L+1}{2}\right\rfloor\right)$ is $M^+\left(L,\left\lfloor\frac{L+1}{2}\right\rfloor\right)$. Denote $M^+\left(L,\left\lfloor\frac{L+1}{2}\right\rfloor\right)$ as M_L^+, we have

$$M_L^+=\left\lfloor\frac{2^{L+1}}{3}\right\rfloor. \tag{4}$$

M_L^+ is also the largest number which can be represented by a canonic SPT interger where the largest power-of-two term is less than or equal to 2^{L-1}.

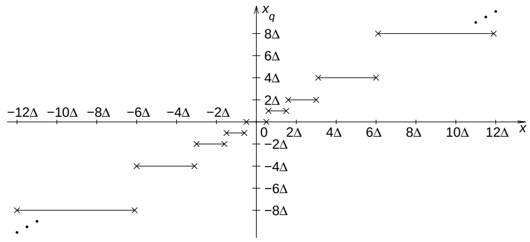

Fig. 1. Input-output characteristics of an SPT roundoff quantizer with $K = 1$, $\Delta = 2^Q = 2^0$.

Let

$$M^+_{L\infty} = \sum_{k=0}^{\infty} 2^{L-1-2k} = \frac{2^{L+1}}{3}. \quad (5)$$

$M^+_{L\infty}$ is the largest infinite precision number represented by SPT terms with the largest power-of-two term being less than or equal to 2^{L-1}.

Furthermore, let $S(L, K)$ be the union of $S^+(L, K)$, $-S^+(L, K)$ and the element 0, i.e., $S(L, K)$ is the integer set that can be represented by L-bit SPT integer with not more than K SPT terms, including both the positive and negative numbers as well as the zero element. Obviously, the number of elements in $S(L, K)$ is $2M^+(L, K) + 1$.

III. Statistics of the SPT Roundoff Error

An SPT roundoff quantizer (to quantize a real number to a sum of K SPT terms, with a precision of 2^Q) is a device having a staircase type input-output relation, as shown in Fig. 1, with a countable number of quantization levels. At each instant, the output of the quantizer can be determined precisely in terms of its input. However, if the input is random, the statistical behavior of the output, rather than this point-by-point response, is of interest. Hence a statistical model is sought to represent the quantization error.

We consider the input to the quantizer as a random variable $x(i), i \in \mathcal{Z}$ and assume that $x(i)$ is uniformly distributed in $[-M^+_{L\infty}, M^+_{L\infty}]$, where $[a, b]$ denotes all the infinite precision numbers in the range bounded by a and b inclusive, as shown in Fig. 2. Denoted as $\mathcal{U}(-M^+_{L\infty}, M^+_{L\infty})$, the probability density function of the random variable $x(i)$ is given by

$$p(x) = \begin{cases} \dfrac{1}{2M^+_{L\infty}}, & -M^+_{L\infty} \le x \le M^+_{L\infty}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The reason for making the above assumption is that the error incurred in a number $x(i)$ to be quantized to SPT form is related to the magnitude of the number. Larger magnitude number may have larger error for a given K. Therefore, the distribution of $x(i)$ affects the distribution of the rounding error.

At each sampling instant i, the quantized output $x_q(i)$, the quantization error $e(i)$, and the input $x(i)$ are related by

$$e(i) = x(i) - x_q(i) \qquad \text{for all } i \in Z^+, \quad (7)$$

where

$$\begin{aligned} x(i), e(i) &\in \mathcal{R}, \\ x_q(i) &\in S(L, K, Q). \end{aligned} \quad (8)$$

In (8), $\mathcal{R}$ is the set of real number, and $S(L, K, Q)$ is the set of SPT number; the element of the set $S(L, K, Q)$ has not more than K SPT terms, and the allowed largest and smallest power of two terms are 2^{L-1} and 2^Q, respectively.

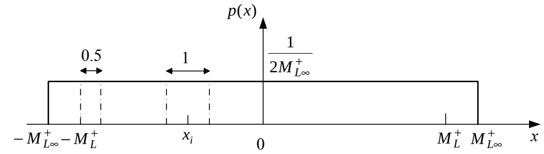

Fig. 2. PDF for a uniformly distributed random number x, $x \in \{x| - M^+_{L\infty} \le x \le M^+_{L\infty}, x \in \mathcal{R}\}$.

For notational convenience, we drop the index i and represent the quantization error model (7) as

$$e = x - x_q. \quad (9)$$

Let $\bar{x}$ be the integer nearest to x.

In this section, we give out 2 properties without proof.

Property 1: The error PDF for rounding a random value x with PDF of $\mathcal{U}(-M^+_{L\infty}, M^+_{L\infty})$ to an element in $S(L, K, Q)$, when $L - Q \ge 2K$ is given by:

$$p_{L,K,Q}(e) = \begin{cases} \dfrac{2M^+(L-Q,K)+1}{2M^+_{L\infty}}, & \text{for } e \in [-2^{Q-1}, 2^{Q-1}], \\ \dfrac{4M^+(L-Q-1-k,K) - 2M^+(L-Q-k,K)+1}{2M^+_{L\infty}}, & \\ & \text{for } e \in \pm[2^{Q+k-1}, 2^{Q+k}], \\ & k = 0, \cdots, L-Q-2K-1, \\ \dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right] \\ & \text{when } L-Q-2K-1 \ge 0, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

□

An example of the error PDF for rounding an infinite precision number to an element in $S(0, 2, -8)$ is shown in Fig.3.

We shorthand the above random distribution of the SPT roundoff error as $\mathcal{E}_u(L, K, Q)$.

Property 2: The mean of the error caused by rounding a random value x with PDF of $\mathcal{U}(-M^+_{L\infty}, M^+_{L\infty})$ to an element in $S(L, K, Q)$, $E(e)$, is equal to 0. The variance of the error, $\sigma^2_{L,K,Q}(e)$, is given by (11).

$$\sigma^2_{L,K,Q}(e) = \begin{cases} \dfrac{2^{2Q}}{12} \cdot \dfrac{27M^+_{L-Q} + 32}{27M^+_{L-Q} + 18}, & \text{for } L - Q = 2K, \\ \dfrac{2^{3Q}}{2M^+_{L\infty}} \Bigg[\left(\dfrac{7}{3}M^+_{2K} + \dfrac{128}{81}\right) 2^{3(L-Q-2K-1)} \\ \quad - \displaystyle\sum_{k=0}^{L-Q-2K-2} 7M^+(L-Q-1-k,K)2^{3k} \\ \quad - M^+(L-Q,K) \Bigg], & \text{for } L - Q \ge 2K + 1. \end{cases} \quad (11)$$

□

Several values of $\sigma_{L,K,Q}(e)$ for L ranges from 4 to -2 and K ranges from 2 to 6, corresponding to $Q = -10$ are listed in Table I. In Table I, each row of values corresponds to the range $\left[-\frac{2^{L+1}}{3}, \frac{2^{L+1}}{3}\right]$. It can be seen from Table I that the variance decreases with increasing K for a given L and decreases with decreasing L for a given K. Therefore, to achieve approximately the same variance, say 10^{-3}, when rounding a number to an SPT value, larger L requires larger

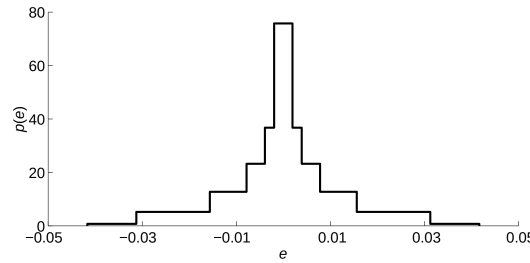

Fig. 3. The error PDF for rounding a number to an SPT number with K SPT terms, and with the allowed largest and smallest power-of-two terms 2^{L-1} and 2^Q, where $L = 0$, $Q = -8$ and K=2.

TABLE I
VALUES OF $\sigma_{L,K,Q}$ FOR $Q = -10$.

L \ K	2	3	4	5	6
4	1.9646E−1	3.6496E−2	6.8449E−3	1.3019E−3	3.4860E−4
3	9.8229E−2	1.8248E−2	3.4244E−3	6.7739E−4	2.9171E−4
2	4.9115E−2	9.1242E−3	1.7178E−3	3.9782E−4	2.8194E−4
1	2.4557E−2	4.5629E−3	8.7432E−4	3.0129E−4	–
0	1.2279E−2	2.2839E−3	4.7644E−4	2.8202E−4	–
−1	6.1396E−3	1.1497E−3	3.1988E−4	–	–
−2	3.0707E−3	5.9718E−4	2.8234E−4	–	–

K, i.e., more SPT terms. When $L = 4$, five SPT terms are required, whereas for $L = -1$, three SPT terms are required. Interestingly, increasing the value of L from -1 to 4 corresponds to an increase in 5 digits in the SPT number but it is only necessary to increase the non-zero digit by 2 digits (i.e., increasing from 3 SPT terms to 5 SPT terms) in order to maintain the same roundoff noise power.

IV. QUANTIZATION OF GAUSSIAN SIGNALS

Gaussian processes are an important class of input models to the quantizer in many practical applications. We study the statistical properties of the quantization error for Gaussian inputs in this section.

A. Approximation of Gaussian Distribution

Consider the input x to be a Gaussian random variable with zero mean and standard deviation σ, denoted as $\mathcal{N}(0, \sigma^2)$. Let the PDF of the input be

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-x^2/(2\sigma^2)). \tag{12}$$

This Gaussian input could be approximated as a weighted superposition of several uniform distributions as

$$\mathcal{N}(0, \sigma^2) \sim \bigcup_{l=L}^{L-2} k_l \times 2M^+_{l\infty}\mathcal{U}(-M^+_{l\infty}, M^+_{l\infty}), \tag{13}$$

where $k \times (b - a)\mathcal{U}(a, b)$ is a weighted uniform distribution with PDF defined as

$$p(x) = \begin{cases} k \times (b-a)\dfrac{1}{b-a} \\ 0 \end{cases} = \begin{cases} k, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Therefore, the Gaussian PDF is approximated as a piece-wise constant function given by

$$p(x) \simeq \begin{cases} k_L, & \text{for } x \in \pm\left[M^+_{L\infty}, M^+_{(L-1)\infty}\right], \\ k_L + k_{L-1}, & \text{for } x \in \pm\left[M^+_{(L-1)\infty}, M^+_{(L-2)\infty}\right], \\ k_L + k_{L-1} + k_{L-2}, & \text{for } x \in \pm\left[M^+_{(L-2)\infty}, 0\right], \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

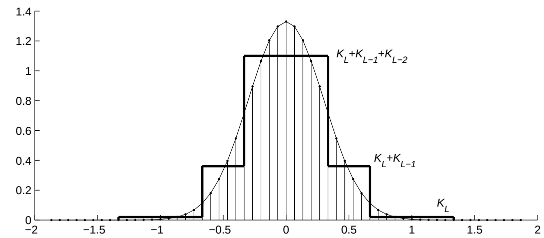

Fig. 4. A Gaussian input $\mathcal{N}(0, 0.3^2)$ is approximated by a piecewise constant distribution.

where

$$L = \overline{\log_2(12\sigma) - 1}, \tag{16}$$

and

$$k_L = \frac{\Phi\left(\frac{-M^+_{(L-1)\infty}}{\sigma}\right)}{M^+_{(L-1)\infty}}, \tag{17}$$

$$k_{L-1} = \frac{\Phi\left(\frac{-M^+_{(L-2)\infty}}{\sigma}\right) - \Phi\left(\frac{-M^+_{(L-1)\infty}}{\sigma}\right)}{M^+_{(L-2)\infty}} - k_L, \tag{18}$$

$$k_{L-2} = \frac{\Phi(0) - \Phi\left(\frac{-M^+_{(L-2)\infty}}{\sigma}\right)}{M^+_{(L-2)\infty}} - k_L - k_{L-1}. \tag{19}$$

Equation (16) is obtained from (5) by supposing that the maximum absolute value of the random variable x is 4σ, and $\overline{\bullet}$ is the integer nearest to $\bullet$. In (17) $-$ (19), $\Phi(x)$ is the Cumulative Distribution Function of Gaussian distribution given by [6]

$$\Phi(x) = \begin{cases} 0.5(1 + \text{erf }(x/\sqrt{2})) & x \ge 0, \\ 0.5(1 - \text{erf }(|x|/\sqrt{2})) & x < 0, \end{cases} \tag{20}$$

where, erf (x) is the error function given by

$$\text{erf }(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{21}$$

For example, an input with $\mathcal{N}(0, 0.3^2)$ can be approximated as a piece-wise constant function as given in (15) and as shown in Fig. 4, where $L = 1$, $k_L = 0.01970, k_{L-1} = 0.34068$ and $k_{L-2} = 0.73984$. Our results show that, as far as round off error analysis is concerned, approximating the Gaussian PDF function by the piecewise constant model of (15) produces insignificant difference. In order to facilitate theoretical analysis of roundoff noise, in this paper, the PDF of a signal will be approximated by a suitable piecewise constant model.

B. Approximation of Gaussian Input Quantization Error Distribution

In this paper, when rounding the input x with $\mathcal{N}(0, \sigma^2)$ distribution, the error analysis is done by approximating it to the error analysis of rounding an input with distribution given in (15). Therefore, the distribution of quantization error (by quantizing the Gaussian input $\mathcal{N}(0, \sigma^2)$ to SPT number with not more than K SPT terms and with the smallest power-of-two term not smaller than 2^Q), shorthanded as $\mathcal{E}_n(\sigma^2, K, Q)$, can be approximated as the weighted superposition of error distributions corresponding to each of the weighted piecewise constant input distribution, i.e.,

$$\mathcal{E}_n(\sigma^2, K, Q) \sim \bigcup_{l=L}^{L-2} k_l \times 2M^+_{l\infty}\mathcal{E}_u(L, K, Q). \tag{22}$$

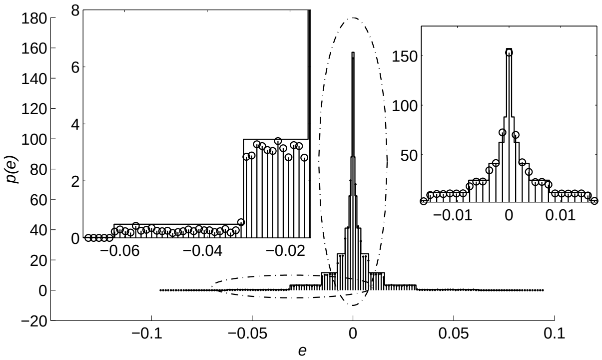

Fig. 5. The distribution of quantization error due to rounding a Gaussian input $\mathcal{N}(0, 0.3^2)$ to SPT numbers with $K = 2$ and $Q = -10$: the solid plots correspond to the piece-wise constant approximated result whereas the histogram plots correspond to the actual measured distribution.

TABLE II
ACTUAL QUANTIZATION ERROR VARIANCES OF 8 SETS OF RANDOM SEQUENCES WITH $\mathcal{N}(0, 0.3^2)$ AND THEIR PIECE-WISE CONSTANT ESTIMATIONS.

No.	Variance	No.	Variance	No.	Variance	No.	Variance
1	1.2644E−4	3	1.2280E−4	5	1.2212E−4	7	1.2421E−4
2	1.2562E−4	4	1.2612E−4	6	1.2556E−4	8	1.2316E−4
Average: 1.2450E−4							
Piece-wise constant estimation: 1.1876E−4							
Relative estimation error: -4.6%							

With the above approximation, the error PDF is approximated as

$$p_{\sigma,K,Q}(e) \simeq \sum_{l=L}^{L-2} k_l \times 2M^{+}_{l\infty} p_{l,K,Q}(e), \tag{23}$$

where, L is given by (16), k_l is given by (17)−(19), and $p_{l,K,Q}(e)$ is given by (10). Correspondingly, the variance of the quantization error is approximated as

$$\sigma^2_{\sigma^2,K,Q}(e) \simeq \sum_{l=L}^{L-2} k_l \times 2M^{+}_{l\infty} \sigma^2_{l,K,Q}(e), \tag{24}$$

where, $\sigma_{l,K,Q}(e)$ is given by (11). The mean value of the Gaussian quantization error remains zero because of the symmetry of the distribution.

C. Examples

The first example illustrates the estimation of the quantization error by rounding x with $\mathcal{N}(0, 0.3^2)$ to SPT numbers with $K = 2$ (i.e. two SPT terms) and $Q = -12$ (i.e. the smallest SPT term is 2^{-12}). The estimation error distribution approximated by (22), as well as the histogram of the quantization errors of a sequence of random numbers and with $\mathcal{N}(0, 0.3^2)$ are plotted in Fig. 5. The sample size of the sequence is 50,000. In the sequel, the same sample size is used for the example sequences. The estimated error variance computed by (24) is 1.1876×10^{-4}, and the actual error variance for the random sequence is 1.2644×10^{-4}. The error variance together with those of other 7 sets of random sequences are listed in Table. II. The relative estimation error in Table. II is defined as

$$\frac{\text{Estimated value} - \text{Average value}}{\text{Average value}}. \tag{25}$$

The second example illustrates the estimation of the quantization errors by rounding Gaussian inputs with $\sigma = 0.2, 0.5$ and 0.8,

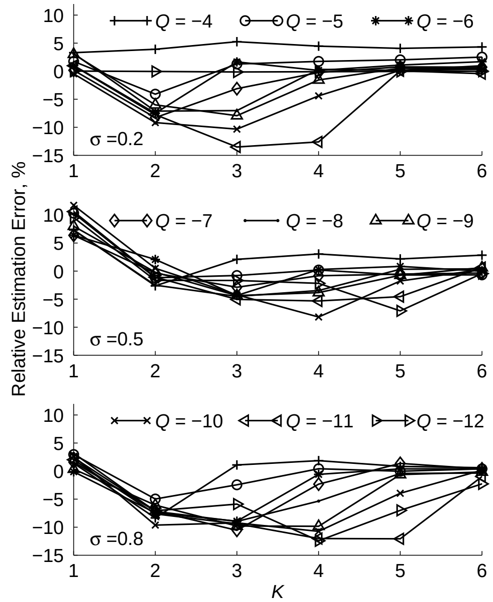

Fig. 6. Relative estimation errors of quantization error variances for rounding Gaussian inputs $\mathcal{N}(0, 0.2^2)$, $\mathcal{N}(0, 0.5^2)$, and $\mathcal{N}(0, 0.8^2)$, respectively, for Q ranging from -4 to -12 and K ranging from 1 to 6.

respectively, to SPT numbers with various K and Q. The relative estimation errors between the estimated error variances and the average actual error variance over 8 trials are plotted in Fig. 6.

As can be seen from Fig. 6, the absolute value of relative estimation error is less than 15% for all cases. For most cases, the relative estimation errors drift from positive values to negative values, and eventually regress to the vicinity of zero when K increases. This is because when K is larger than half of the wordlength, the SPT roundoff is the same as the uniformly distributed fix-point roundoff.

V. CONCLUSION

In this paper, the error distribution for quantizing infinite precision signals to SPT values for a given L, K, Q is deduced, where K is the number of SPT terms, 2^{L-1} and 2^Q are the allowed largest and smallest power-of-two terms, respectively, given the input signals are uniformly distribute in $[-M^{+}_{L\infty}, M^{+}_{L\infty}]$. The roundoff errors for Gaussian distributed signals are obtained by approximating the signals as a weighted superposition of several uniformly distributed signals. Simulation results show that such piece-wise constant model faithfully represents the Gaussian model in roundoff error variance analysis producing a maximum relative estimation error of less than 15%. This translates to $10 * \log_{10}(1.15) = 0.61$dB in error power.

REFERENCES

[1] Y. C. Lim, J. B. Evans, and B. Liu, “Decomposition of binary integers into signed power-of-two terms,” *IEEE Trans. Circuits Syst.*, vol. 38, pp. 667–672, June 1991.
[2] Y. C. Lim, R. Yang, D. N. Li, and J. J. Song, “Signed power-of-two term allocation scheme for the design of digital filters,” *IEEE Trans. Circuits Syst. II*, vol. 46, pp. 577–584, May 1999.
[3] A. B. Sripad and D. L. Snyder, “A necessary and sufficient condition for quantization errors to be uniform and white,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 442–448, Oct. 1977.
[4] C. Y. Yao, “A study of SPT-term distribution of csd numbers and its application for designing fixed-point linear phase FIR filter,” in *Proc. IEEE ISCAS’01*, vol. II, 2001, pp. 301–304.
[5] J. J. Bussgang, “Crosscorrelation funtions of amplitude-distorted Gaussian signals,” Research laboratory of electronics, MIT, Cambridge, Massachusetts, Tech. Rep. 26, Mar. 1952.
[6] X. R. Li, *Probability, random signals and statistics*. CRC Press, 1999.