Roundoff noise analysis of signals represented using signed power-of-two terms

Yu, Ya Jun; Lim, Yong Ching

2007

Yu, Y. J., & Lim, Y. C. (2007). Roundoff noise analysis of signals represented using signed power-of-two terms. *IEEE Transactions on Signal Processing*, 55(5). 2122-2135.

<https://hdl.handle.net/10356/91619>

<https://doi.org/10.1109/TSP.2007.893216>

IEEE Transactions on Signal Processing © 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.
<http://www.ieee.org/portal/site>.

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

Ya Jun Yu, *Member, IEEE*, and Yong Ching Lim, *Fellow, IEEE*

Abstract—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. The design of digital filters with power-of-two 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 sum of a limited number of signed power-of-two (SPT) terms is available, 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 and so 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. The result is useful for determining the number of SPT terms required to represent a signal subject to a given roundoff noise.

Index Terms—Gaussian model, probability density function, roundoff error, signed power-of-two (SPT), statistical analysis.

I. INTRODUCTION

THE implementation of digital filters with digital devices having finite wordlength unavoidably introduces quantization errors. There are three main sources of quantization error that can arise: input quantization [1], [2], coefficient quantization [3]–[6], and quantization in arithmetic operations [7]. Extensive studies on the statistics of finite wordlength roundoff errors have been widely reported in the literature, from statistical models [2], [8], to effects on various filter structures, such as direct form FIR filters [3], cascade FIR filters [9], [10], and recursive FIR filters [11].

There has been increasing interests on the representation of signals and digital filter coefficient values as a limited number of signed power-of-two (SPT) terms [12]–[18], during the past two decades, for the fact that multiplications involving SPT numbers in binary arithmetic can be replaced by add/subtract-shift operations.

It is well known that when a number is represented as a sum of SPT terms, the number of nonzero digits is less than or equal to that when it is represented in two's complement. For canonic

SPT representation, no two consecutive nonzero digits exist. Therefore, an R -bit number can be represented using no more than $(R + 1)/2$ SPT terms. Often, fewer terms are needed, and it has been shown in [19] that the expected number of SPT terms in an R -bit canonic SPT number tends asymptotically to $((R/3) + (1/9))$ as R increases. Digital filters whose coefficient values are integer power-of-two [12]–[18], [20]–[23] are essentially multiplierless. However, the design of digital filters with 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 [24], 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 [12] and so 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 [2], 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 [12] and [25]. Lim *et al.* [12] 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 [25] derived the distribution of the SPT terms, i.e., the probability of allocating an SPT term to a particular digit. 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. The result is useful for determining the number of SPT terms required to represent a signal to within an acceptable roundoff noise level. Section II reviews some properties of SPT numbers; these properties will be used in the subsequent derivations. An implementation for an adaptive filter when input signals are SPT numbers is described in Section III. In Section IV, 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 V, an estimation of the roundoff error for Gaussian distributed inputs is approximated. An example illustrating the application of our result is presented in Section VI, where the signal is a clip from the song "Memory" sung by Barbara Streisand.

Manuscript received June 18, 2005; revised June 8, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soren Holdt Jendou. This work was supported in part by Nanyang Technological University (NTU), Singapore, and in part by Temasek Laboratory, NTU.

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798 Singapore (e-mail: eleyuyj@pmail.ntu.edu.sg; eleyimyc@pmail.ntu.edu.sg).

Digital Object Identifier 10.1109/TSP.2007.893216

II. SPT 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

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

where $Q \leq q(i) \leq L-1$. $L-Q$ is a positive integer but Q may be zero, a positive integer, or a negative integers. Furthermore, for any i and j , it satisfies the constraints that $q(i) \neq q(j)$ if $i \neq j$ and that $q(i) \neq q(j) + 1$. The canonic representation requirement imposes the further constraint [i.e., $q(i) \neq q(j) + 1$ in (1)] on the representation; this will exclude the representation of some numbers that can be represented without the canonic representation requirement under the same wordlength. For example, with $L = 4$ and $Q = 0$ in (1), one cannot represent say 12 in the canonic form but it is possible in the noncanonic form. To represent 12 in a canonic form, it is necessary to increase L to 5 which is a drawback compared to the case with $L = 4$ that allows noncanonic forms. However, because of the unique representation of the canonic SPT number which is very attractive for monitoring and ensuring the minimum representation of the number, many researches imposed the canonic condition on the SPT numbers for easy analysis and derivation [12], [22], [25]–[27], although there may exist other minimum representations. For the same reasons, in this paper, the canonic condition is implied in the discussions.

For the particular condition in (1) where $Q = 0$, n is an L -digit integer. In this paper, this particular case for quantizing a number to an SPT number where $Q = 0$ is considered first. The case where $Q \neq 0$ will be deduced from the case where $Q = 0$.

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 number has at most $\lfloor (L+1)/2 \rfloor$ SPT terms (or nonzero digits), where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x . Assume further that $L \geq 2K - 1$. 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 [12]

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

where $K!$ 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

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

Fig. 1. Adaptive filter.

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

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

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

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. EXAMPLE OF APPLICATION

The computation of each output sample $y(n)$ at time n for the filter shown in Fig. 1 requires N multiplications.

$$y(n) = \mathbf{X}(n)^T \mathbf{W}(n) \quad (6)$$

where

$$\mathbf{X}(n) = [x(n), x(n-1), \dots, x(n-N+1)]^T \quad (7)$$

and

$$\mathbf{W}(n) = [w_1(n), w_2(n), \dots, w_N(n)]^T. \quad (8)$$

In binary arithmetic, the multiplication process is essentially multiplierless if either the elements of $\mathbf{X}(n)$ or $\mathbf{W}(n)$ are SPT numbers. For fix filters, either the elements of $\mathbf{X}(n)$ or $\mathbf{W}(n)$ may be made SPT since it is possible to optimize $\mathbf{W}(n)$ in the SPT space using sophisticated optimization techniques [12]–[18], [20]–[23] during the design process. For adaptive filters, it is impractical to optimize $\mathbf{W}(n)$ in the SPT space. Thus, in the implementation of adaptive filters, it is necessary

Fig. 2. Implementation structure for the filter shown in Fig. 1, when the input signals are represented as SPT numbers.

to convert the elements of $\mathbf{X}(n)$ to SPT numbers if a multiplierless implementation is desired. This can be achieved by preceding the filter with a low complexity SPT converter [24], [28] converting $x(n)$ into SPT numbers. When the elements of $\mathbf{X}(n)$ are in SPT format, the computation of the vector inner product $\mathbf{X}(n)^T \mathbf{W}(n)$ can be achieved in a multiplierless fashion using the structure as shown in Fig. 2.

When the elements of $\mathbf{X}(n)$ are in SPT format, the arithmetic complexity (which in turn influences power consumption, clock speed, and silicon area requirement) of the filter depends on the number of SPT terms and the wordlength of $x(n)$. In order to reduce the arithmetic complexity to the minimum, it is necessary to reduce the number of SPT terms and wordlength for $x(n)$. Nevertheless, reducing the SPT terms and wordlength will result in an increase in the roundoff noise for $x(n)$. This paper presents an analysis for the roundoff noise introduced into $x(n)$ when $x(n)$ is rounded to a limited number of SPT terms subject to a given wordlength.

IV. STATISTICS OF THE SPT ROUND OFF 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. 3, 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. 4. Denoted as $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$, the probability density function of the random variable $x(i)$ is given by

$$p(x) = \begin{cases} \frac{1}{2M_{L\infty}^+}, & -M_{L\infty}^+ \leq x \leq M_{L\infty}^+ \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Fig. 3. Input-output characteristic of an SPT roundoff quantizer with $K = 1$, $\Delta = 2^Q$.

Fig. 4. PDF for a uniformly distributed random number x , $x \in \{x | -M_{L\infty}^+ \leq x \leq M_{L\infty}^+, x \in \mathcal{R}\}$.

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

$$e(i) = x(i) - x_q(i) \text{ for all } i \in \mathcal{Z}^+ \quad (10)$$

where

$$\begin{aligned} x(i), e(i) &\in \mathcal{R} \text{ (the set of real numbers)} \\ x_q(i) &\in S(L, K). \end{aligned} \quad (11)$$

For notational convenience, we drop the index i and represent the quantization-error model (10) as

$$e = x - x_q. \quad (12)$$

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

A. Error Probability Density Function

Lemma 1: Given a random value x with PDF of $\mathcal{U}(-M_L^+, M_L^+)$, and given $\bar{x} \in S(L, K)$, the PDF of the roundoff error caused by quantizing x to the nearest element in $S(L, k)$ is $\mathcal{U}(-1/2, 1/2)$. $\square$

The proof for Lemma 1 is shown in Appendix A.

1) *Property 1:* For $L = 2K - 1$, the error PDF of rounding a random value x With PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to the nearest element in $S(L, K)$, denoted as $p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \frac{2M_{L\infty}^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in [-3^{-1}, 3^{-1}] \\ \frac{M_{L\infty}^+}{M_{L\infty}^+}, & \text{for } e \in \pm [3^{-1}, 2^{-1}] \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

□

Proof: For $L = 2K - 1$, according to (4) and (5), it is obvious that $M_{L\infty}^+ = M_L^+ + 3^{-1}$. For any value $x \in \{x | -M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$, we have $\bar{x} \in S(L, K)$. According to Lemma 1, the rounding error is uniformly distributed in $[-2^{-1}, 2^{-1}]$ with the unity PDF. Furthermore, the probability that $x \in \{x | -M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ is $2M_L^+ / (2M_{L\infty}^+)$. The probability that $x \in \{x | M_L^+ \leq x \leq M_{L\infty}^+\}$ and $x \in \{x | -M_{L\infty}^+ \leq x \leq -M_L^+\}$ are both $1/(6M_{L\infty}^+)$, and the rounding errors are distributed uniformly in $[0, 3^{-1}]$ and $[-3^{-1}, 0]$, respectively, with PDF equal to 3. Therefore

$$p_{L,K}(e) = \begin{cases} \frac{2M_L^+}{2M_{L\infty}^+} \times 1, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{1}{6M_{L\infty}^+} \times 3, & \text{for } e \in [0, 3^{-1}] \\ \frac{1}{6M_{L\infty}^+} \times 3, & \text{for } e \in [-3^{-1}, 0] \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{2M_L^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in [-3^{-1}, 3^{-1}] \\ \frac{M_L^+}{M_{L\infty}^+}, & \text{for } e \in \pm [3^{-1}, 2^{-1}] \\ 0, & \text{otherwise.} \end{cases}$$

■

2) *Property 2:* For $L = 2K$, the error PDF for rounding a random value x with PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to the nearest member in $S(L, K)$, denoted as $p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \frac{2M_{L\infty}^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in [-\frac{1}{2}, \frac{1}{2}] \\ \frac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm [\frac{1}{2}, \frac{2}{3}] \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

□

The proof for property 2 is similar to that for property 1.

Lemma 2: Given a random value x with PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$, and given $\bar{x} \in T(L, K + 1)$, according to (1), $\bar{x}$ can be written as $\bar{x} = \sum_{i=0}^K y(i)2^{q(i)}$. Given the least significant SPT term of $\bar{x}$ is $y(K)2^{q(K)}$, where $q(K) = L - 2K - 1 \geq 0$ and $y(K) = 1$ or -1 , the error caused by rounding x to the nearest member in $S(L, K)$ is uniformly distributed with unity PDF in $[2^{q(K)} - 2^{-1}, 2^{q(K)}] \cup [-2^{q(K)}, -2^{q(K)} + 2^{-1}]$. □

The proof for Lemma 2 is shown in Appendix B.

Lemma 3: For $L = 2K + 1$, the error PDF for rounding a random value x With PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to an element in $S(L, K)$, denoted as $p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \frac{2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{2N^+(L,K+1)-1}{2M_{L\infty}^+}, & \text{for } e \in \pm [2^{-1}, 2^0] \\ \frac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm [2^0, \frac{4}{3}] \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

□

The proof for Lemma 3 is shown in Appendix C.

Lemma 4: For $L > 2K + 1$, the error PDF for rounding a random value x with PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to an element in $S(L, K)$, denoted as $p_{L,K}(e)$, is given in (16), shown at the bottom of the page. The proof for Lemma 4 is shown in Appendix D.

3) *Property 3:* The error PDF for rounding a random value x with PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to an element in $S(L, K)$ when $L \geq 2K$ is given by (17), shown at the bottom of the next page. □

Proof: According to property 2, it is noted that (17) is true for $L = 2K$. According to Lemma 3, (15) is true when $L = 2K + 1$. Furthermore, according to (3), when $L = 2K + 1$, $M^+(L, K + 1) = 2M^+(L - 1, K) + 1$. Since $M^+(L, K + 1) = M^+(L, K) + N^+(L, K + 1)$, therefore

$$2N^+(L, K + 1) - 1 = 4M^+(L - 1, K) - 2M^+(L, K) + 1. \quad (18)$$

Substituting (18) into (15), we have

$$p_{L,K}(e | e \in \pm [2^{-1}, 2^0]) = \frac{4M^+(L - 1, K) - 2M^+(L, K) + 1}{2M_{L\infty}^+}. \quad (19)$$

$$p_{L,K}(e) = \begin{cases} \frac{2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{4M^+(L-1,K)-2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in \pm [2^{-1}, 2^0] \\ \frac{p_{L-1,K}(e | e \in \pm [2^{k-1}, 2^k])}{2}, & \text{for } e \in \pm [2^k, 2^{k+1}], k = 0, \dots, L - 2k - 2 \\ \frac{p_{L-1,K}(e | e \in \pm [2^{l-2k-2}, \frac{2^{l-2k}}{3}])}{2}, & \text{for } e \in \pm [2^{l-2k-1}, \frac{2^{l-2k+1}}{3}] \text{ when } l - 2K - 1 \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

Therefore, (15) can be written as

$$p_{L,K}(e) = \begin{cases} \frac{2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{4M^+(L-1,K)-2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in \pm[2^{-1}, 2^0] \\ \frac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm[2^0, \frac{2^2}{3}] \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Hence, (17) is true for $L = 2K + 1$.

For $L > 2K + 1$, (17) is proved by mathematical induction as follows. Assume that the error PDF for rounding x to an element in $S(L, K)$ is given by (17). For some particular value of L , where the PDF of x is $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$. According to Lemma 4 and the fact that $2M_{L\infty}^+ = M_{(L+1)\infty}^+$, we obtain (21), shown at the bottom of the page. Hence, if (17) is true for L , it is also true for $L + 1$. Since (17) is true for $L = 2K + 1$, it is true for all integer $L \geq 2K + 1$.

Therefore, the error probability density function for $L \geq 2K$ can be written as (17). And, thus, property 3 is proved. ■

So far, we have obtained the error PDF for rounding a random value x with PDF of $\mathcal{U}(-M_{L\infty}^+, M_{L\infty}^+)$ to an SPT integer (i.e., the smallest power-of-two term is 2^0) with not more than K SPT terms. An example of the error PDF for rounding an infinite precision number to an element in $S(8, 2)$ is shown in Fig. 5.

If the smallest power-of-two term is 2^Q instead of 2^0 , where the largest power-of-two term remains at 2^{L-1} , the word length of the resulting SPT number becomes $L-Q$; the error PDF of this quantization process is a scaled version of $p_{L-Q,K}(e)$, where the value of the error PDF is scaled by $1/2^Q$ and the error range is scaled by 2^Q . Therefore, when $L - Q \geq 2K$, the error PDF for quantizing a number to an SPT number with K SPT terms, and with the largest and smallest power-of-two terms 2^{L-1} and 2^Q , respectively, denoted as $p_{L,K,Q}(e)$, is

Fig. 5. Error PDF for rounding a number to an L digit SPT integer with not more than K SPT terms, where $L = 8$ and $K = 2$.

given by (22), shown at the bottom of the next page. We short hand the above random distribution of the SPT roundoff error as $\mathcal{E}_u(L, K, Q)$.

B. Mean and Variance

1) *Property 4:* 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 SPT number with k SPT terms, and with the largest and smallest power-of-two terms 2^{L-1} and 2^Q , respectively, $E(e)$ is equal to 0. The variance of the error $\sigma_{L,K,Q}^2(e)$ is given by (23), shown at the bottom of the next page. □

This property can be proved in a straight forward manner.

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, where each row of values corresponds to the range $[-2^{L+1}/3, 2^{L+1}/3]$. 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 K , i.e., more SPT terms. When $L = 4$, five SPT terms are required, whereas for $L = -1$,

$$p_{L,K}(e) = \begin{cases} \frac{2M^+(L,K)+1}{2M_{L\infty}^+}, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{4M^+(L-1-k,K)-2M^+(L-k,K)+1}{2M_{L\infty}^+}, & \text{for } e \in \pm[2^{k-1}, 2^k], k = 0, \dots, L-2K-1 \\ \frac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm[2^{L-2K-1}, \frac{2^{L-2K+1}}{3}] \text{ when } L-2K-1 \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

$$p_{L+1,K}(e) = \begin{cases} \frac{2M^+(L+1,K)+1}{2M_{(L+1)\infty}^+}, & \text{for } e \in [-2^{-1}, 2^{-1}] \\ \frac{4M^+(L+1-1-k,K)-2M^+(L+1-k,K)+1}{2M_{(L+1)\infty}^+}, & \text{for } e \in \pm[2^{k-1}, 2^k], k = 0, \dots, L+1-2K-1 \\ \frac{1}{2M_{(L+1)\infty}^+}, & \text{for } e \in \pm[2^{L+1-2K-1}, \frac{2^{L+1-2K+1}}{3}] \text{ when } L+1-2K-1 \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

Fig. 6. $\sigma_{L,K,Q}(e)$ plot for $L = 5$. (a) $K = 2$. (b) $K = 3$. Note that the error variance decreases with decreasing Q for a given number range $[-2^{L+1}/3, 2^{L+1}/3]$ and a given K .

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 nonzero digit by 2 digits (i.e., increasing from 3 SPT terms to 5 SPT terms) in order to maintain the same roundoff noise power.

If the largest power-of-two term is fixed, i.e., if L is fixed, the error variance for representing a random number in SPT form decreases with decreasing Q for a given K . In Fig. 6(a), the largest power-of-two term is fixed at $2^{L-1} = 2^4$, thus, the number representable is in $[-64/3, 64/3]$, and $K = 2$. When

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	2.8189E-4
0		1.2279E-2	2.2839E-3	4.7644E-4	2.8202E-4	—
-1		6.1396E-3	1.1497E-3	3.1988E-4	2.8183E-4	—
-2		3.0707E-3	5.9718E-4	2.8234E-4	—	—

Q decreases from 0 to -3 , i.e., the smallest power-of-two term decreases from 2^0 to 2^{-3} , the error variance drops quickly from 0.431 to 0.393. However, for Q , the decrease in the variance is almost not noticeable and the variance approaches a constant asymptotically as Q decreases. Increasing K may reduce the variance further, as shown in Fig. 6(b) for $K = 3$. However, the variance in Fig. 6(b) approaches asymptotically a smaller constant as Q decreases.

V. 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

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (24).$$

$$\begin{aligned}
 P_{L,K,Q}(e) &= P_{L-Q,K}(2^{-Q}e) \cdot \frac{1}{2^Q} \\
 &= \begin{cases} \frac{2M^+(L-Q,K)+1}{2M_{L_\infty}^+}, & \text{for } e \in [-2^{Q-1}, 2^{Q-1}] \\ \frac{4M^+(L-Q-1-k,K)-2M^+(L-Q-k,K)+1}{2M_{L_\infty}^+}, & \text{for } e \in \pm[2^{Q+k-1}, 2^{Q+k}], k = 0, \dots, L-Q-2K-1 \\ \frac{1}{2M_{L_\infty}^+}, & \text{for } e \in \pm\left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3}\right] \text{ when } L-Q-2K-1 \geq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (22)
 \end{aligned}$$

$$\sigma_{L,K,Q}^2(e) = \begin{cases} \frac{2^{2Q}}{12} \cdot \frac{27M_{L-Q}^+ + 4}{27M_{L-Q}^+ + 9}, & \text{for } L-Q = 2K-1 \\ \frac{2^{2Q}}{12} \cdot \frac{27M_{L-Q}^+ + 32}{27M_{L-Q}^+ + 18}, & \text{for } L-Q = 2K \\ \frac{2^{3Q}}{2M_{L_\infty}^+} \left[\left(\frac{7}{3}M_{2K}^+ + \frac{128}{81} \right) 2^{3(L-Q-2K-1)} - \sum_{k=0}^{L-Q-2K-2} 7M^+(L-Q-1-k,K)2^{3k} - M^+(L-Q,K) \right], & \text{for } L-Q \geq 2K+1. \end{cases} \quad (23)$$

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

$$\mathcal{N}(0, \sigma^2) \sim \bigcup_{l=L}^{L-2} k_l \times 2M_{l\infty}^+ \mathcal{U}(-M_{l\infty}^+, M_{l\infty}^+) \quad (25)$$

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

$$p(x) = \begin{cases} k, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases} \quad (26)$$

Therefore, the Gaussian PDF is approximated as a piece-wise constant function given by (27), shown at the bottom of the page, where

$$L = \overline{\log_2(12\sigma) - 1} \quad (28)$$

and

$$k_L = \frac{\Phi\left(\frac{-M_{(L-1)\infty}^+}{\sigma}\right)}{M_{(L-1)\infty}^+} \quad (29)$$

$$k_{L-1} = \frac{\Phi\left(\frac{-M_{(L-2)\infty}^+}{\sigma}\right) - \Phi\left(\frac{-M_{(L-1)\infty}^+}{\sigma}\right)}{M_{(L-2)\infty}^+} - k_L \quad (30)$$

$$k_{L-2} = \frac{\Phi(0) - \Phi\left(\frac{-M_{(L-2)\infty}^+}{\sigma}\right)}{M_{(L-2)\infty}^+} - k_L - k_{L-1}. \quad (31)$$

Equation (28) is obtained from (5) by supposing that the maximum absolute value of the random variable x is 4σ , and $\bullet$ is the integer nearest to $\bullet$. In (29)–(31), $\Phi(x)$ is the cumulative distribution function of Gaussian distribution given by [29]

$$\Phi(x) = \begin{cases} 0.5 \left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right) & x \geq 0 \\ 0.5 \left(1 - \operatorname{erf}\left(\frac{|x|}{\sqrt{2}}\right)\right) & x < 0 \end{cases} \quad (32)$$

where $\operatorname{erf}(x)$ is the error function given by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (33)$$

For example, an input with $\mathcal{N}(0, 0.3^2)$ can be approximated as a piece-wise constant function as given in (27) and as shown in Fig. 7, 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

Fig. 7. Gaussian input $\mathcal{N}(0, 0.3^2)$ is approximated by a piecewise constant distribution.

function by the piecewise constant model of (27) 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 (27). 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), shorthand 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.,

$$\mathcal{E}_n(\sigma^2, K, Q) \sim \bigcup_{l=L}^{L-2} k_l \times 2M_{l\infty}^+ \mathcal{E}_u(L, K, Q). \quad (34)$$

With the above approximation, the error PDF is approximated as

$$p_{\sigma^2, K, Q}(e) \simeq \sum_{l=L}^{L-2} k_l \times 2M_{l\infty}^+ p_{l, K, Q}(e) \quad (35)$$

where, L is given by (28), k_l is given by (29)–(31), and $p_{l, K, Q}(e)$ is given by (22). Correspondingly, the variance of the quantization error is approximated as

$$\sigma_{\sigma^2, K, Q}^2(e) \simeq \sum_{l=L}^{L-2} k_l \times 2M_{l\infty}^+ \sigma_{l, K, Q}^2(e) \quad (36)$$

$$p(x) \simeq \begin{cases} k_L, & \text{for } x \in \pm [M_{L\infty}^+, M_{(L-1)\infty}^+] \\ k_L + k_{L-1}, & \text{for } x \in \pm [M_{(L-1)\infty}^+, M_{(L-2)\infty}^+] \\ k_L + k_{L-1} + k_{L-2}, & \text{for } x \in \pm [M_{(L-2)\infty}^+, 0] \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

Fig. 8. Distribution of quantization error due to rounding a Gaussian input $\mathcal{N}(0, 0.3^2)$ to SPT numbers with $K = 2$ and $Q = -12$. 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

n	Variance	n	Variance	n	Variance	n	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%							

where, $\sigma_{1,K,Q}(e)$ is given by (23). 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 (34), as well as the histogram of the quantization errors of a sequence of random numbers with $\mathcal{N}(0, 0.3^2)$ are plotted in Fig. 8. The estimated error variance computed by (36) 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 seven sets of random sequences are listed in Table II. The relative estimation error in Table II is defined as

$$\frac{\text{Estimated value} - \text{Average value}}{\text{Average value}}. \quad (37)$$

The second example illustrates the estimation of the quantization errors by rounding Gaussian inputs with $\sigma = 0.2, 0.5, 0.8$ to SPT numbers with various K and Q . The estimated error variances and average actual variances over eight trials are listed in Table III.

It can be seen from Table III that the estimations of the SPT quantization error variances are close to the actual ones, especially when the number of power-of-two term (K value) is small. The estimation deviates a little further from the actual one when K increases. However, as can be seen from Fig. 9,

TABLE III
ESTIMATED AND ACTUAL QUANTIZATION ERROR VARIANCES FOR
RANDOM SEQUENCES WITH $\mathcal{N}(0, 0.2^2)$, $\mathcal{N}(0, 0.5^2)$ AND
 $\mathcal{N}(0, 0.8^2)$, FOR VARIOUS K AND Q VALUES

	Estimated	Actual Average	Relative Error
$\sigma = 0.2$			
$K = 1, Q = -12$	1.5446E-3	1.5315E-3	0.9%
$K = 2, Q = -14$	5.1214E-5	5.4860E-5	-6.6%
$K = 3, Q = -16$	1.7674E-6	1.9509E-6	-9.4%
$\sigma = 0.5$			
$K = 2, Q = -12$	3.4004E-4	3.4130E-4	-0.4%
$K = 3, Q = -14$	1.1735E-5	1.2227E-5	-4.0%
$K = 4, Q = -16$	4.1269E-7	4.3176E-7	-4.4%
$\sigma = 0.8$			
$K = 3, Q = -12$	2.8278E-5	3.1292E-5	-9.6%
$K = 4, Q = -14$	9.9453E-7	1.1228E-6	-11.4%
$K = 5, Q = -16$	3.5280E-8	4.0187E-8	-12.2%

Fig. 9. Relative estimation errors of quantization error variances for rounding a Gaussian input $\mathcal{N}(0, 1^2)$, for Q ranging from -10 to -16 and K ranging from 1 to 10.

the relative estimation error is less than 15% for all cases. In Fig. 9, 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.

VI. EXAMPLE WHERE THE SIGNAL IS A PIECE OF MUSIC

The PDFs of many signals that exist in real life are not strictly Gaussian but are modeled as Gaussian for the purpose of theoretical analysis. Following the same argument, our analysis and approximation technique can also be used to estimate the roundoff error power and signal-to-noise ratio (SNR) of real life signals that are not strictly Gaussian. Our technique can thus be used to determine the number of SPT terms required to represent a real life signal so that the roundoff noise is acceptable.

In this section, we shall choose a clip from the song "Memory" sung by Barbara Streisand. The signal has a variance of 0.006894 and its distribution resemble that of Gaussian with the same variance, as shown in Fig. 10.

TABLE IV

ESTIMATED AND ACTUAL SNR FOR ROUNDING THE SIGNAL OF A PIECE OF MUSIC TO SPT NUMBER WITH VARIOUS K AND Q VALUES. THE VALUES IN THE SQUARE BRACKETS ARE ESTIMATED SNR AND THE VALUES IN THE ROUND BRACKETS ARE THE CORRESPONDING ACTUAL SNR. THE PAIRS OF Q AND K VALUES WHICH ARE OVERLY CONSERVATIVE ARE DARKENED

$K \backslash Q$	1	2	3	4	5	6
-2	[21.35] (22.79)	--	--	--	--	--
-3	[26.92] (28.29)	[26.92] (28.29)	--	--	--	--
-4	[32.92] (32.99)	[33.82] (34.08)	--	--	--	--
-5	[34.48] (34.74)	[39.71] (40.00)	[39.71] (40.00)	--	--	--
-6	[34.72] (35.07)	[45.53] (45.38)	[45.98] (46.00)	--	--	--
-7	[34.74] (35.12)	[48.57] (48.41)	[51.97] (52.03)	[51.97] (52.03)	--	--
-8	[34.75] (35.13)	[49.38] (49.28)	[57.83] (57.72)	[58.05] (58.04)	--	--
-9	[34.75] (35.13)	[49.52] (49.54)	[62.01] (61.82)	[64.06] (64.11)	[64.06] (64.11)	--
-10	[34.75] (35.13)	[49.54] (49.46)	[63.69] (63.46)	[69.99] (69.92)	[70.10] (70.09)	--
-11	[34.75] (35.13)	[49.55] (49.47)	[64.08] (63.86)	[74.95] (74.76)	[76.12] (76.11)	[76.12] (76.12)
-12	[34.75] (35.13)	[49.55] (49.47)	[64.15] (63.93)	[77.65] (77.35)	[82.09] (81.90)	[82.14] (81.98)
-13	[34.75] (35.13)	[49.55] (49.47)	[64.17] (63.95)	[78.49] (78.16)	[87.52] (86.89)	[88.16] (87.59)
-14	[34.75] (35.13)	[49.55] (49.47)	[64.17] (63.94)	[78.67] (78.34)	[91.21] (89.97)	[94.16] (92.40)
-15	[34.75] (35.13)	[49.55] (49.47)	[64.17] (63.95)	[78.70] (78.41)	[92.72] (92.59)	[99.86] (107.75)

Fig. 10. Histogram of a piece of music and a Gaussian fit with an identical variance.

Tabulated in Table IV are the estimated and actual SNR for rounding the signals with various K and Q values. In Table IV, the values in square brackets are the estimated SNR and the values in round brackets are the corresponding actual SNR. As can be seen in Table IV, the errors between the estimated SNR and the actual SNR are very small. It can also be seen in Table IV that the SNR increases asymptotically with increasing K and/or with decreasing Q to certain values. The increase in SNR is neglectable when increasing K or decreasing Q beyond these values.

In Table IV, darkened squares correspond to pairs of Q and K values that are overly conservative, i.e., Q can be increased and K can be decreased without deteriorating the SNR value by more than 1 dB. Consider for example that darkened square where $Q = -8$ and $K = 4$; the SNR value, as indicated in Table IV is 58.05 dB. Reducing K to 3 will cause the SNR value to deteriorate to 57.83 dB; the difference is only 0.22 dB. Consider another darkened square where $Q = -11$ and $K = 3$; the SNR is 64.08. Increasing Q to -10 will cause only a 0.39-dB deterioration in SNR. Thus, ideally the pair of Q and K selected should correspond to an undarkened square in Table IV.

For example, to round signals with SNR 30, 50, 70, and 90 dB, the recommended (K, Q) pairs are $(1, -4)$, $(3, -7)$, $(4, -11)$, and $(5, -14)$, respectively.

The squares in Table IV are darkened based upon our technique. If the darkened squares are obtained using actual experimental data, the results are very close to that obtained in Table IV. In this particular example, the only difference is the square corresponding to $Q = -4$ and $K = 2$; in Table IV, it is a darkened square but if it is obtained based upon experimental results, it is a white square.

VII. 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 largest and smallest power-of-two terms, respectively, given the input signals are uniformly distributed 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. We illustrate, by means of an example, that, although our technique is developed based upon a Gaussian model, it produces sufficiently accurate roundoff noise estimation for a real life signal which has close resemblance to Gaussian but is not strictly Gaussian. Our technique is valuable for estimating the number of SPT terms and wordlength required to meet a particular SNR.

APPENDIX A PROOF OF LEMMA 1

Proof: For a random value x with PDF of $\mathcal{U}(-M_L^+, M_L^+)$, since $\bar{x} \in S(L, K)$, if $\bar{x} \neq \pm M_L^+$, the rounding error PDF is unity in $[-2^{-1}, 2^{-1}]$ [2]. The probability of $\bar{x} \neq \pm M_L^+$ given

that $\bar{x} \in S(L, K)$ is $(2M_L^+ - 1)/2M_L^+$. The probabilities of $\bar{x} = M_L^+$ or $\bar{x} = -M_L^+$ given that $\bar{x} \in S(L, K)$ are $0.5/2M_L^+$, respectively. The rounding error PDF in each of $[0, 0.5]$ and $[-0.5, 0]$ is 2. Denote the error PDF as $p_{L,K}(e, x | -2^{-1} \leq e \leq 2^{-1}, \bar{x} \in S(L, K))$, we have

$$p_{L,K}(e, x | -2^{-1} \leq e \leq 2^{-1}, \bar{x} \in S(L, K)) = \begin{cases} \frac{2M_L^+ - 1}{2M_L^+} \times 1, & \text{for } -2^{-1} \leq e \leq 2^{-1} \\ \frac{0.5}{2M_L^+} \times 2, & \text{for } 0 \leq e \leq 2^{-1} \\ \frac{0.5}{2M_L^+} \times 2, & \text{for } -2^{-1} \leq e \leq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (38)$$

APPENDIX B PROOF OF LEMMA 2

Proof: According to the given conditions, $\bar{x}$ is an L -bit integer with $K + 1$ SPT terms; these $K + 1$ SPT terms are occurring at the most significant $2K + 1$ bits. Since in canonic representations, there are no two nonzero adjacent bits, the $K + 1$ bits corresponding to the $K + 1$ SPT terms must occur at every other bit locations.

First, consider the case where x is positive. We have $\bar{x} - 2^{-1} \leq x < \bar{x} + 2^{-1}$ and

$$\bar{x} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K-1}. \quad (39)$$

From (39), it can be shown that

$$\bar{x} - y(K)2^{L-2K-1} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} \in S(L, K) \quad (40)$$

and that

$$\bar{x} + y(K)2^{L-2K-1} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K}. \quad (41)$$

When $\bar{x} \neq \sum_{i=0}^{K-1} y(i)2^{L-2i-1}$, at least one of $y(i)$ for $i = 0, 1, \dots, K-1$ is not equal to $y(K)$, i.e., they have different signs. An example of $\bar{x}$ is shown in Fig. 11(a), where

■ Fig. 11. For $\bar{x} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K-1}$, we have $\bar{x} + y(K)2^{L-2K-1} \in S(L, K)$.

$y(K) = 1$ and $y(K-2)$ has a different sign from $y(K)$. Thus, $\bar{x} + y(K)2^{L-2K-1}$ has $K + 1$ SPT terms as shown in Fig. 11(c), where two nonzero bits occur adjacently. From Fig. 11(c)–(e), it is shown that this number (with $K + 1$ SPT terms), $\bar{x} + y(K)2^{L-2K-1}$ as shown in Fig. 11(c), can be represented using K SPT terms canonically, as shown in Fig. 11(e), provided that there exists an integer $i > 0$ and for all j where $0 \leq j < i$ such that $y(K-i) = -y(K)$ and $y(K-j) = y(K)$. In this example, i equals to 2. Therefore, under the given conditions, $\bar{x} + y(K)2^{L-2K-1}$ is an element in $S(L, K)$, i.e.,

$$\bar{x} + y(K)2^{L-2K-1} \in S(L, K). \quad (42)$$

From (40) and (42), it is shown that

$$\bar{x} \pm y(K)2^{L-2K-1} \in S(L, K). \quad (43)$$

Moreover, it is obvious that $\bar{x} \pm 2^{L-2K-j} \notin S(L, K)$ for any $j > 1$. That means any number in $(\bar{x} - 2^{L-2K-1}, \bar{x} + 2^{L-2K-1})$ is not in $S(L, K)$, where (a, b) denotes all the infinite precision numbers in the range bounded by a and b exclusive. Therefore, the result of rounding x to an element in $S(L, K)$ is equal to $\bar{x} \pm 2^{q(K)}$.

$$p_{L,K}(e, x | e \in \pm [2^{q(K)} - 2^{-1}, 2^{q(K)}], \bar{x} \in T(L, K+1), q(K) = L - 2K - 1) = \begin{cases} \frac{2N^+(L, K+1)-1}{2J} \times 1 & \text{for } e \in \pm [2^{q(K)} - 2^{-1}, 2^{q(K)}] \\ \frac{J-(2N^+(L, K+1)-1)}{2J} \times 2 & \text{for } e \in [2^{q(K)} - 2^{-1}, 2^{q(K)}] \\ \frac{J-(2N^+(L, K+1)-1)}{2J} \times 2 & \text{for } e \in -[2^{q(K)} - 2^{-1}, 2^{q(K)}] \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{for } e \in \pm [2^{q(K)} - 2^{-1}, 2^{q(K)}] \\ 0 & \text{otherwise.} \end{cases} \quad (44)$$

Since x is uniformly distributed in $[\bar{x} - 2^{-1}, \bar{x} + 2^{-1}]$, so that rounding x to the element in $S(L, K)$ nearest to it will lead to an error distributed uniformly in $\pm [2^{q(K)} - 2^{-1}, 2^{q(K)}]$, with PDF equal to $1/(2[2^{q(K)} - (2^{q(K)} - 2^{-1})]) = 1$. For a negative number x and $\bar{x} \neq -\sum_{i=0}^K 2^{L-2i-1}$, the same conclusion can be obtained. The probability of $\bar{x} \neq \pm \sum_{i=0}^K 2^{L-2i-1}$ for the given $\bar{x}$ is $(2N^+(L, K+1) - 1)/J$, where $J = 2N^+(L, K+1) + 1$ if $\sum_{i=0}^K 2^{L-2i-1} + 0.5 \leq M_{L\infty}^+$; otherwise, $J = 2N^+(L, K+1) + 2(M_{L\infty}^+ - \sum_{i=0}^K 2^{L-2i-1})$.

When $\bar{x} = \sum_{i=0}^K 2^{L-2i-1}$ (occurring with probability of $(J - (2N^+(L, K+1) - 1))/2J$), x is uniformly distributed in $[\bar{x} - 2^{-1}, \bar{x}]$. Furthermore, $\bar{x} - 2^{L-2K-1} \in S(L, K)$ and $\bar{x} - 2^{L-2K-j} \notin S(L, K)$ for any $j > 1$. Therefore, the error is uniformly distributed in $[2^{q(K)} - 2^{-1}, 2^{q(K)}]$, with PDF equal to $1/(2^{q(K)} - (2^{q(K)} - 2^{-1})) = 2$. Similarly, when $\bar{x} = -\sum_{i=0}^K 2^{L-2i-1}$ (occurring with probability of $(J - (2N^+(L, K+1) - 1))/2J$), the rounding error is uniformly distributed in $[-2^{q(K)}, -2^{q(K)} + 2^{-1}]$ with PDF equal to 2. Therefore, we arrive at (44), shown at the bottom of the previous page.

APPENDIX C PROOF OF LEMMA 3

Proof: First, $x \in \{x | -M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ is considered. In this case, since $L = 2K + 1$, we have $M_L^+ = M^+(L, K+1)$ and $\bar{x} \in S(L, K) \cup T(L, K+1)$. The probability for $x \in \{x | -M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ and $\bar{x} \in S(L, K)$ is $(2M^+(L, K) + 1)/2M_{L\infty}^+$ and is denoted as $P_L(x | \bar{x} \in S(L, K))$. According to Lemma 1, rounding x for $\bar{x} \in S(L, K)$ to an element in $S(L, K)$ leads to an error in $[-2^{-1}, 2^{-1}]$. Therefore

$$\begin{aligned} p_{L,K}(e | e \in [-2^{-1}, 2^{-1}]) \\ = P_L(x | \bar{x} \in S(L, K)) \times p_{L,K}(e, x | e \in [-2^{-1}, 2^{-1}], \\ \bar{x} \in S(L, K)) \\ = \frac{2M^+(L, K) + 1}{2M_{L\infty}^+}. \end{aligned} \quad (45)$$

The probability for $x \in \{x | -M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ and $\bar{x} \in T(L, K+1)$, denoted as $P_L(x | \bar{x} \in T(L, K+1))$, is $(2N^+(L, K+1) - 1)/2M_{L\infty}^+$. For $L = 2K + 1$ and $\bar{x} \in T(L, K+1)$, we have $\bar{x} = \sum_{i=0}^K y(i)2^{q(i)}$ where $q(K) = 0$. According to Lemma 2, rounding x for $\bar{x} \in T(L, K+1)$ to an element in $S(L, K)$ nearest to it leads to an error distributed in $\pm [2^{-1}, 2^0]$. Therefore, the error PDF for $e \in \pm [2^{-1}, 2^0]$ is

$$\begin{aligned} p_{L,K}(e | e \in \pm [2^{-1}, 2^0]) &= P_L(x | \bar{x} \in T(L, K+1)) \\ &\quad \times p_{L,K}(e, x | e \in \pm [2^{-1}, 2^0], \\ &\quad \bar{x} \in T(L, K+1)) \\ &= \frac{2N^+(L, K+1) - 1}{2M_{L\infty}^+}. \end{aligned} \quad (46)$$

Finally, the probabilities for $x \in \{x | M_L^+ \leq x \leq M_{L\infty}^+, x \in \mathcal{R}\}$ and $x \in \{x | -M_{L\infty}^+ \leq x \leq -M_L^+, x \in \mathcal{R}\}$ are both $(M_{L\infty}^+ - M_L^+)/2M_{L\infty}^+$; rounding x to an element in $S(L, K)$ leads to an error which is uniformly distributed in $\pm [2^0, \sum_{k=0}^{\infty} 2^{-2k}] = \pm [1, 4/3]$ with PDF equal to 3. Therefore

$$\begin{aligned} p_{L,K}(e | e \in \pm [1, \frac{4}{3}]) \\ = \begin{cases} \frac{M_{L\infty}^+ - M_L^+}{2M_{L\infty}^+} \times 3, & \text{for } e \in [1, \frac{4}{3}] \\ \frac{M_{L\infty}^+ - M_L^+}{2M_{L\infty}^+} \times 3, & \text{for } e \in [-\frac{4}{3}, -1] \end{cases} \\ = \frac{1}{2M_{L\infty}^+}. \end{aligned} \quad (47)$$

Combining (45), (46), and (47), (15) is proved. $\blacksquare$

APPENDIX D PROOF OF LEMMA 4

Proof: Assume that when $L \geq 2K + 1$, the error PDF for rounding a number x' , $x' \in \{x' | -M_{L\infty}^+ \leq x' \leq M_{L\infty}^+, x' \in \mathcal{R}\}$ to an element in $S(L, K)$ is shown in (48), shown at the bottom of the page. In (48), the error $e \in [-2^{-1}, 2^{-1}]$ is produced when the integer part of x' has not more than K SPT

$$\begin{aligned} p_{L,K}(e) \\ = \begin{cases} p_{L,K}(e | e \in [-2^{-1}, 2^{-1}]), & \text{for } e \in [-2^{-1}, 2^{-1}] \\ p_{L,K}(e | e \in \pm [2^{k-1}, 2^k]), & \text{for } e \in \pm [2^{k-1}, 2^k], k = 0, \dots, L - 2K - 1 \\ p_{L,K}(e | e \in \pm [2^{L-2K-1}, \frac{2^{L-2K+1}}{3}]), & \text{for } e \in \pm [2^{L-2K-1}, \frac{2^{L-2K+1}}{3}] \text{ when } L - 2K - 1 \geq 0 \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (48)$$

terms, whereas the errors in the other ranges in (48) are produced when the integer part of x' has more than K SPT terms. From Lemma 3, it is shown that this assumption is true for $L = 2K + 1$.

Since $M_{(L+1)\infty}^+ = 2M_{L\infty}^+$, the number $x = 2x'$ is obtained by shifting the binary point for x' by one bit to the right.

- 1) When the integer part of x' has more than K SPT terms, as shown in Fig. 12(a), it is obvious that the error caused by rounding $x = 2x'$ to an element in $S(L+1, K)$ is distributed in $\pm[2^k, 2^{k+1}]$ for $k = 0, \dots, L - 2K - 1$, or an error in $\pm[2^{L-2K}, 2^{L-2K+2}/3]$. Therefore

$$\begin{aligned} p_{L+1,K} \left(e \middle| e \in \pm[2^k, 2^{k+1}] \right) &\times 2(2^{k+1} - 2^k) \\ &= p_{L,K} \left(e \middle| e \in \pm[2^{k-1}, 2^k] \right) \times 2(2^k - 2^{k-1}) \\ &\text{for } k = 0, \dots, L - 2K - 1 \end{aligned} \quad (49)$$

and

$$\begin{aligned} p_{L+1,K} \left(e \middle| e \in \pm \left[2^{L-2K}, \frac{2^{L-2K+2}}{3} \right] \right) &\times \frac{2^{L-2K}}{3} \\ &= p_{L,K} \left(e \middle| e \in \pm \left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3} \right] \right) \times \frac{2^{L-2K-1}}{3}, \\ &\text{for } L - 2K - 1 \geq 0 \end{aligned} \quad (50)$$

where

$$p_{L+1,K} \left(e \middle| e \in \pm[2^k, 2^{k+1}] \right) \quad (51)$$

and

$$p_{L+1,K} \left(e \middle| e \in \pm \left[2^{L-2K}, \frac{2^{L-2K+2}}{3} \right] \right) \quad (52)$$

are the error PDFs for e in $\pm[2^k, 2^{k+1}]$ and $\pm[2^{L-2K}, 2^{L-2K+2}/3]$, respectively, when rounding x to an element in $S(L+1, K)$. From (49) and (50), we have

$$\begin{aligned} p_{L+1,K} \left(e \middle| e \in \pm[2^k, 2^{k+1}] \right) \\ &= \frac{1}{2} \times p_{L,K} \left(e \middle| e \in \pm[2^{k-1}, 2^k] \right) \end{aligned} \quad (53)$$

and

$$\begin{aligned} p_{L+1,K} \left(e \middle| e \in \pm \left[2^{(L+1)-2K-1}, \frac{2^{(L+1)-2K+1}}{3} \right] \right) \\ &= \frac{1}{2} \times p_{L,K} \left(e \middle| e \in \pm \left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3} \right] \right). \end{aligned} \quad (54)$$

- 2) When the integer part of x' has not more than K SPT terms, rounding x' to an element in $S(L, K)$ leads to an error

Fig. 12. Number x' , $x' \in \{x' | -M_L^+ \leq x' \leq M_L^+, x' \in \mathcal{R}\}$ is represented in SPT form. (a) Integer part of x' has more than K SPT terms. (b) Integer part of x' has less than or equal to K SPT terms, where $K = 2$.

uniformly distributed in $[-2^{-1}, 2^{-1}]$ and occurring with probability

$$P_L \left(x' \middle| \bar{x}' \in S(L, K) \right) = \frac{2M^+(L, K) + 1}{2M_{L\infty}^+}. \quad (55)$$

Therefore, rounding $2x'$ to an element in $S(L+1, K)$ leads to an error in $[-2^0, 2^0]$.

For all $\bar{x}' \in S(L, K)$, some of the $\bar{2x}'$ belong to $S(L+1, K)$, whereas the others belong to $S(L+1, K+1)$. For example, the number $\bar{2x}'$, where $\bar{x}'$ is shown in Fig. 12(b), belongs to $S(L+1, K+1)$. For $L+1 > 2K+1$

$$M_{(L+1)\infty}^+ - M^+(L+1, K) \geq \sum_{k=K}^{\infty} 2^{L-2k} > 1. \quad (56)$$

Therefore, the probability of $\bar{x} = \bar{2x}' \in S(L+1, K)$, i.e., $\bar{x}$ can be represented exactly by an element in $S(L+1, K)$, is $(2M^+(L+1, K) + 1)/2M_{(L+1)\infty}^+$ and is denoted as $P_{L+1} \left(x \middle| \bar{x} \in S(L+1, K) \right)$, so that such x will lead to errors in $[-2^{-1}, 2^{-1}]$. The error PDF for $e \in [-2^{-1}, 2^{-1}]$ is thus given by

$$\begin{aligned} p_{L+1,K}(e | -2^{-1} \leq e \leq 2^{-1}) \\ &= P_{L+1} \left(x \middle| \bar{x} \in S(L+1, K) \right) \\ &\quad \times p_{L+1,K} \left(e, x \middle| -2^{-1} \leq e \leq 2^{-1}, \right. \\ &\quad \left. \bar{x} \in S(L+1, K) \right) \end{aligned} \quad (57)$$

where $p_{L+1,K} \left(e, x \middle| -2^{-1} \leq e \leq 2^{-1}, \bar{x} \in S(L+1, K) \right)$ is equal to 1 according to Lemma 1. Therefore, we have

$$\begin{aligned} p_{L+1,K}(e | -2^{-1} \leq e \leq 2^{-1}) &= \frac{2M^+(L+1, K) + 1}{2M_{(L+1)\infty}^+} \times 1 \\ &= \frac{2M^+(L+1, K) + 1}{2M_{(L+1)\infty}^+}. \end{aligned} \quad (58)$$

Rounding the remaining x , where $\bar{x} \in S(L+1, K+1)$, will lead to errors distributed in $\pm[2^{-1}, 2^0]$. We have

$$p_{L+1,K}(e|e \in \pm[2^{-1}, 2^0]) \\ = \left[P_L \left(x' \middle| \bar{x}' \in S(L, K) \right) \right. \\ \left. - P_{L+1} \left(x \middle| \bar{x} \in S(L+1, K) \right) \right] \\ \times p_{L+1,K} \left(e, x \middle| x \in \pm[2^{-1}, 2^0], \right. \\ \left. \bar{x} \in S(L+1, K+1) \right)$$

where $p_{L+1,K}(e, x| x \in \pm[2^{-1}, 2^0], \bar{x} \in S(L+1, K+1))$ is equal to 1 according to Lemma 4. Therefore, we have

$$p_{L+1,K}(e|e \in \pm[2^{-1}, 2^0]) \\ = \left(\frac{2M^+(L, K) + 1}{2M_{L\infty}^+} - \frac{2M^+(L+1, K) + 1}{2M_{(L+1)\infty}^+} \right) \times 1 \\ = \frac{4M^+(L, K) - 2M^+(L+1, K) + 1}{2M_{(L+1)\infty}^+}. \quad (59)$$

Combining (58), (59), (53), and (54), it can be seen that if the assumption in (48) is true for $p_{L,K}(e)$ for a particular value of L , (16) is true for $p_{L+1,K}(e)$. As the assumption in (48) is true for $p_{L,K}(e)$ where $L = 2K + 1$ as stated in Lemma 3, thus it is also true for $p_{L+1,K}(e)$. Hence, (4) is true for $p_{L,K}(e)$ for all $L \geq 2K + 1$. ■

REFERENCES

- [1] S. K. Mitra, K. Hirano, and H. Sakaguchi, "A simple method of computing the input quantization and multiplication roundoff errors in a digital filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-22, no. 10, pp. 326–329, Oct. 1974.
- [2] 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 Process.*, vol. ASSP-25, no. 10, pp. 442–448, Oct. 1977.
- [3] J. B. Knowles and E. M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuit Theory*, vol. CT-15, no. 1, pp. 31–41, Mar. 1968.
- [4] D. S. K. Chan and L. R. Rabiner, "Analysis of quantization errors in the direct form for finite impulse response digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 2, pp. 354–366, Aug. 1973.
- [5] R. E. Crochiere, "A new statistical approach to the coefficient word length problem for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, no. 2, pp. 190–196, Mar. 1975.
- [6] J. J. Nielsen, "Design of linear-phase direct-form digital filters with quantized coefficients using error spectrum shaping techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 1020–1026, Jul. 1989.
- [7] J. P. Dugre, A. A. L. Beex, and L. L. Scharf, "Generating co variance sequences and the calculation of quantization and rounding error variances in digital filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 102–104, Feb. 1980.
- [8] C. W. Barnes, B. N. Tran, and S. H. Leung, "On the statistics of fixed-point roundoff error," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 3, pp. 595–606, Jun. 1985.
- [9] D. S. K. Chan and L. R. Rabiner, "Theory of roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 229–245, Mar. 1973.
- [10] D. S. K. Chan and L. R. Rabiner, "An algorithm for minimizing roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 347–385, Mar. 1973.
- [11] C. Caraiscos and B. Liu, "A roundoff error analysis of the LMS adaptive algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 1, pp. 34–41, Feb. 1984.
- [12] 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, Analog Digit. Signal Process.*, vol. 46, no. 5, pp. 577–584, May 1999.
- [13] D. N. Li, Y. C. Lim, Y. Lian, and J. J. Song, "A polynomial-time algorithm for design digital filters with power-of-two coefficients," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1935–1941, Aug. 2002.
- [14] Y. C. Lim and S. R. Parker, "FIR filter design over a discrete power-of-two coefficient space," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 6, pp. 583–591, Jun. 1983.
- [15] Y. C. Lim and Y. J. Yu, "A width-recursive depth-first tree search approach for the design of discrete coefficient perfect reconstruction lattice filter bank," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 50, no. 6, pp. 257–266, Jun. 2003.
- [16] C.-L. Chen and A. N. Willson, Jr., "A trellis search algorithm for the design of FIR filters with signed-powers-of-two coefficients," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 46, no. 1, pp. 29–39, Jan. 1999.
- [17] E. Eweda, "Convergence analysis and design of an adaptive filter with finite-bit power-of-two quantized error," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 39, no. 2, pp. 113–115, Feb. 1992.
- [18] Y. J. Yu and Y. C. Lim, "A novel genetic algorithm for the design of a signed power-of-two coefficient quadrature mirror filter lattice filter bank," *Circuit Syst. Signal Process.*, vol. 21, pp. 263–276, May/Jun. 2002.
- [19] G. W. Reitwiesner, "Binary arithmetic," *Advances in Computers*, vol. 1, pp. 231–308, 1960.
- [20] Y. C. Lim and S. R. Parker, "Discrete coefficient FIR digital filter design based upon an LMS criteria," *IEEE Trans. Circuits Syst.*, vol. CAS-30, no. 5, pp. 723–739, Oct. 1983.
- [21] Y. C. Lim, "Design of discrete-coefficient-value linear phase FIR filters with optimum normalized peak ripple magnitude," *IEEE Trans. Circuits Syst.*, vol. 37, no. 12, pp. 1480–1486, Dec. 1990.
- [22] B.-R. Horng, H. Samuelli, and A. N. Willson, Jr., "The design of two-channel lattice structure perfect-reconstruction filter banks using power-of-two coefficients," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, pp. 497–499, Jul. 1993.
- [23] Y. J. Yu, "Multiplierless multirate FIR filter design and implementation," Ph.D. dissertation, Elect. Comp. Eng. Dept., National Univ. Singapore, Singapore, 2003.
- [24] 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, no. 6, pp. 667–672, Jun. 1991.
- [25] 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*, 2001, vol. II, pp. 301–304.
- [26] H. Samuelli, "An improved search algorithm for the design of multiplierless FIR filters with powers-of-two coefficients," *IEEE Trans. Circuits Syst.*, vol. 36, no. 7, pp. 1044–1047, Jul. 1989.
- [27] A. Fuller, B. Nowrouzian, and F. Ashrafzadeh, "Optimization of FIR digital filters over the canonical signed-digit coefficient space using genetic algorithms," in *Proc. Midwest Symp. Circuits Syst.*, Aug. 1998, pp. 456–459.
- [28] E. Backenius, E. Säll, and O. Gustafsson, "Bidirectional conversion to minimum signed-digit representation," in *Proc. IEEE ISCAS'06*, 2006, pp. 2413–2416.
- [29] X. R. Li, *Probability, Random Signals and Statistics*. Boca Raton, FL: CRC Press, 1999.

Ya Jun Yu (S'99–M'05) received the B.Sc. and M.Eng. degrees in biomedical engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2004.

From 1997 to 1998, she was a Teaching Assistant with Zhejiang University. From 1998 to 2004, she was a Research Engineer in the Department of Electrical and Computer Engineering, National University of Singapore. In 2002, she was a Visiting

Researcher at the Tampere University of Technology, Tampere, Finland, and the Hong Kong Polytechnic University, Hong Kong. From 2004 to 2005, she was a Research Fellow in the Temasek Laboratories, Nanyang Technological University, Singapore. Since 2005, she has been with School of Electrical and

Electronic Engineering, Nanyang Technological University as an Assistant Professor. Her research interests include digital signal processing, VLSI circuits, and systems design.

Yong Ching Lim (S'79–M'82–SM'92–F'00) received the A.C.G.I. and B.Sc. degrees in 1977 and the D.I.C. and Ph.D. degrees in 1980 from Imperial College, University of London, U.K., all in electrical engineering.

From 1980 to 1982, he was a National Research Council Research Associate in the Naval Postgraduate School, Monterey, CA. From 1982 to 2003, he was with the Department of Electrical Engineering, National University of Singapore, Singapore. Since 2003, he has been with the School of Electrical and

Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor. His research interests include digital signal processing and VLSI circuits and systems design.

Dr. Lim received the 1996 IEEE Circuits and Systems Society's Guillemin–Cauer Award, the 1990 IREE (Australia) Norman Hayes Memorial Award, 1977 IEE (U.K.) Prize, and the 1974–1977 Siemens Memorial (Imperial College) Award. He was a Lecturer for the IEEE Circuits and Systems Society under the Distinguished Lecturer program from 2001 to 2002 and as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1991 to 1993 and from 1999 to 2001. He has also served as an Associate Editor for *Circuits, Systems and Signal Processing* from 1993 to 2000. He served as the Chairman of the Digital Signal Processing (DSP) Technical Committee of the IEEE Circuits and Systems Society from 1998 to 2000. He served in the Technical Program Committee's DSP Track as the Chairman in IEEE ISCAS'97 and IEEE ISCAS'00 and as a Co-chairman in IEEE ISCAS'99. He is the General Chairman for IEEE APCCAS'06. He is a member of Eta Kappa Nu.