

Study of Additive Dither on Restraining Signal Truncation Error

Tao Liu*, Shulin Tian, Zhigang Wang, Lianping Guo

School of Automation Engineering, University of Electronic Science and Technology of China
Xiyuan Ave, West Hi-Tech Zone, Chengdu, P.R.China, 86-28-61831318

*Corresponding author, e-mail: liutaoprivate@gmail.com

Abstract

Owing to the limited calculation precision during digital signal processing, the intermediate stages' signal-bit-width truncation should be executed to realize the conversion from high precision to low one. As method of direct truncation will degenerate the Spurious FreeDynamicRange (SFDR) performance of the output signal, this paper proposed that additional digital dither should be added before operation of truncation, which could decline the harmonic distortion efficiently and extend the dynamic range of the truncated signal significantly. Two simulations for truncation operation towards signal with additive Gaussian dither and uniform dither are carried out to prove the validity of the proposed method. Comparative studies demonstrate that the proposed algorithm applied in Gaussian dithering and uniform dithering could improve the output SFDR performance by about 16dB and 15dB respectively.

Keywords: digital dither, truncation error, precision conversion, SFDR

Copyright © 2014 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction

With the introduction of advanced Field Programmable Gate Array (FPGA) architectures which provide built-in Digital Signal Processing (DSP) support such as embedded multipliers, block RAMs DSP blocks and so on, a new hardware alternative is available for DSP designers who can get even higher level of performance than those achievable on general purpose DSP processors [1]. In order to reduce the cost of hardware while increasing throughput rates, most digital FPGA implementations of signal algorithms rely solely on fixed-point approximations, such as frequency mixing, signal decimation, and filtering and so on. The inevitable problem caused by fix-point calculation is the growth in bit width. Especially after multistage calculation, the increment of bit width of final output data, compared to that of original input data, turns out to be very considerable. When the latter stages' allowable processing bit width tends to be short, we need to transform the previous stages' output bit width into some extent in order to accommodate the width requirement of the latter stage.

The strategies for bit-width conversion can be roughly categorized into two groups [2]. The first one is basically an analytical approach coming from those algorithm designers who analyze the finite word length effects due to mantissa processing arithmetic. The other approach is based on bit-true simulation origination from the hardware designers. The analytical approach started from attempts to model quantization error statistically; then it was expanded to specific linear time invariant (LTI) systems such as digital filters, FFT, etc. In the past three decades, numerous papers have been devoted to this approach [2-5]. The bit-true simulation method has been extensively used recently [6-8]. Its potential benefits lie in its ability to handle non-LTI systems as well as LTI systems. Whatever, the aforementioned two approaches have the similarity of handling and analyzing data influence at the output port of some certain stage. While this paper proposes a new way, which tends to firstly introduce some difference (i.e. digital dither the below mentioned, compared to analog dither in ADC-optimizing field [9]) into the input port of some certain processing unit and then execute mantissa processing at the output port. There are mainly two types of mantissa processing: truncation and round to nearest. Round to nearest employs an extra adder for the rounding operation, while truncation directly chops the bits lower than required lest significant bit, which is the main type we concern in this paper. Mantissa processing towards multi-bit data, is a transformation process from high quantized precision to low one. Due to the reduction in quantization width, truncation error is

aroused, which leads to the noise peak in the output spectrum caused by harmonic distortion, and reduces the spurious free dynamic range (SFDR).

Sampling operation of ADC can also be considered as a transformation process from high quantized precision to low precision. Dither technology [10-14] aims to reduce the quantization effect of analog-to-digital convertor and improve the SFDR performance drastically, mainly by means of adding random noise into ADC input signal. Based on the dithering principle, this paper mainly draws digital dither into analysis of truncation error and conduct the study of restraining truncation error. In order to improve the harmonic distortion caused by bit-width truncation, we need to add appropriate random noise into the pre-truncated signal with long bit-width before truncation operation is executed. The proposed method works in pure digital domain, which will be easily and efficiently applied to most of the digital processing systems.

2. Truncation Error Analysis

Generally speaking, error sources in digital systems mainly come from the following two aspects: 1) Quantization error exported by ADC; 2) Truncation error introduced by finite word length effect. For options in which the circuit structure is fixed, it is impractical to reduce ADC's quantization error. As a result, truncation error becomes the major target needing resolved.

Digital signal processing is based on a series of algorithm, whose computational accuracy determines the accuracy of the final output result. In order to obtain higher precision in field programmable gate array (FPGA), the bit-width allocated in intermediate stages tends to be longer than that in the final result. For instance, most of digital signal calculation in FPGA are developed in fixed-point arithmetic, which means the operands are all integers. The whole calculation process always contains a series of steps, such as multiplication, filtering, signal compressing and so on, which will absolutely increase the efficient bit-width of intermediate results. Take a look at the logic resource of FFT IP core within ALTERA series' FPGA, whose longest acceptable input bit-width is 24-bits. After operation such as quantization, mixing, filtering and so on, signal, whose bit-width is mostly longer than 24-bits, is sent to the input port of the IP core. Then, redundant bits should be cut so as to satisfy the requirement.

Assume $x(n)$ is the final output signal with effective bit width of A after a series of calculation, and $y(n)$ is the truncated signal with effective bit width of B. Executing fix-pointed calculation in FPGA, the most simple and also the easiest method for truncation is discarding the lowest certain bits directly. Let $z(n)$ represents the discarded error signal with bit width of A-B, then, we get the following equation:

$$x(n) = y(n) \cdot 2^{A-B} + z(n) \quad (1)$$

Let $X(e^{j\omega})$, $Y(e^{j\omega})$ and $Z(e^{j\omega})$ represent the Discrete Fourier Transformation (DFT) of $x(n)$, $y(n)$ and $z(n)$ respectively. According to the linear property of Fourier Transform, we can get:

$$X(e^{j\omega}) = 2^{A-B} \sum_{n=0}^{\infty} y(n)e^{-j\omega n} + \sum_{n=0}^{\infty} z(n)e^{-j\omega n} \quad (2)$$

Which means,

$$X(e^{j\omega}) = 2^{A-B} Y(e^{j\omega}) + Z(e^{j\omega}) \quad (3)$$

Equation (3) indicates that, in the case of relevant spectrum parameters (e.g. SFDR) of $x(n)$ under certain, the spectrum quality of truncated signal $y(n)$ is affected by truncation error $z(n)$.

As mentioned above, $y(n)$ is achieved by discarding the lowest $A-B$ bits of $x(n)$. Signal constructed by the discarded $A-B$ bits is called $z(n)$, then, $x(n)$ and $z(n)$ become established as:

$$z(n) = x(n) \pmod{2^{A-B}} \quad (4)$$

As shown in Equation (4), symbol 'mod' represents the congruence operator, which means the remainder achieved by $x(n)$ dividing 2^{A-B} . Assume $x(n) = m \in [0, 2^{A-1}]$, then Equation (4) is equivalent with the following formula:

$$z(m) = m \pmod{2^{A-B}} \quad (5)$$

Equation (5) means retaining the lowest $A-B$ bits of m . Let $\Delta = 2^{A-B} - 1$, then Equation (5) can be described as Figure 1.

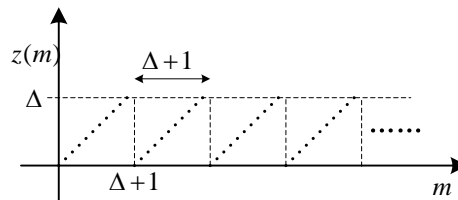


Figure 1. Function of Truncation Error

Having a close-up view of Figure 1, it is not difficult to find that, the truncation error function $z(m)$ has the similar property with ADC's quantization error, especially the periodicity and linearity. The difference lies in that, $z(m)$ is a periodic discrete linear function, with a period of $N = \Delta + 1 = 2^{A-B}$. Such periodicity of the truncation error $z(m)$ is reflected in the harmonic on the output spectrum, which will degenerate the spectrum quality after truncation.

3. Dither Principle

Dither is a kind of random jitter signal, which is completely independent with ADC's analog input signal, [15]. Harmonic may come from ADC's coherent sampling, quantized noise and periodicity of differential nonlinear error, who are created by the existed certain relevance among sampling, quantization and input waveform, [16]. Dither is just employed in order to damage the relative fixed relationship. Adding dither into ADC's input signal and wiping off the noise at the output port with digital methods, the SFDR will be improved. The operation principle is as shown in Figure 2. From the view of amplitude, the additive dither can be classified into dither with high amplitude and low amplitude. While from the view of frequency, the additive dither can be classified into wideband dither and narrowband dither. Application occasions vary with different kinds of dither.

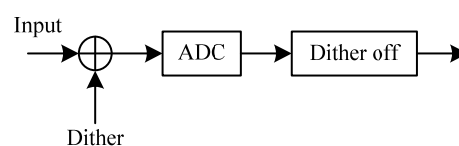


Figure 2. Principle of Dither

The principle of dither shown in Figure 2 involves ADC and its relevant analog circuit design, which is not that easy to implement in a fixed digital system. That's why we need a kind of pure digital method to restrain truncation error, i.e. this paper referred digital dither.

4. Dither-based Restraints of Truncation Error

Similar with ADC's quantization principle, truncation towards long data bit-width is a transformation process from high quantization precision to low precision. During the process, detailed information between two adjacent sample point decrease with the reduction of quantization steps, which leads to the fact that continuously variable details among several sample point of the high precision signal turn into a flat step without any variation. Meanwhile, harmonic distortion is introduced and SFDR is decreased, [17, 18].

Among the range of $x(n) = m \in [0, 2^A - 1]$, truncation error $z(m)$ is a discrete sawtooth wave function with period of N . When $m \in [0, N - 1]$, $z(m) = m$. Thus the N -point Fourier Series of $z(m)$ is listed as:

$$z(m) = \sum_{k=0}^{N-1} a_k e^{jk\omega_0 m} = \sum_{k=0}^{N-1} a_k e^{jk(2\pi/N)m} \quad (6)$$

Coefficient a_k in the Fourier Series expression Equation (6) is:

$$\begin{aligned} a_k &= \frac{1}{N} \sum_{m=0}^{N-1} z(m) e^{-jk\omega_0 m} \\ &= \frac{1}{N} \sum_{m=0}^{N-1} m e^{-jk(2\pi/N)m} \\ &= \frac{\cos(2\pi k / N) - 1 + j \sin(2\pi k / N)}{2[1 - \cos(2\pi k / N)]} \\ &= \frac{j e^{j\pi k / N}}{2 \sin(\pi k / N)} \end{aligned} \quad (7)$$

Fourier transformation expression of periodical signal $z(m)$ within one period can be obtained from Equation (7).

$$Z(e^{j\omega}) = j\pi \sum_{k=0}^{N-1} \frac{e^{j\pi k / N}}{\sin(\pi k / N)} \delta(\omega - \frac{2\pi k}{N}) \quad (8)$$

When k value is small, the amplitude-frequency characteristic function can be described as:

$$|Z(e^{j\omega})| = \sum_k \frac{N}{k} \delta(\omega - \frac{2\pi k}{N}) \quad (9)$$

What can be deduced from Equation (9) is that, variation of harmonic amplitude has a inverse ratio to the value of k , and attenuates slowly. In order to rapidly attenuate the harmonic amplitude in the truncation error, we plan to add random noise with width of $A - B$ bits into signal $x(n)$ before truncation operation is executed. The added random noise is the so-called dither discussed in this paper. Width of dither is based on the following consideration: 1) Over-high amplitude will introduce floor-noise-rising of $y(n)$; 2) Undersize amplitude is inadequate

for changing the step-characteristic of $y(n)$. The additive dither with $A - B$ bits should not affect the floor-noise, but will well change the step-characteristic of $y(n)$.

As dither is a statistic signal, we need to represent truncation error with mathematical expectation. Let signal d indicates the additive dither, and $d \in [0, 2^{A-B} - 1] \cap Z$, where Z represents the set of integers, which means d is a discrete random variable. Assume the new truncation error is $z'(m)$, then:

$$\begin{aligned} z'(m) &= E[z(m + d)] \\ &= \sum_{l=0}^{N-1} z(m + l) p(l) \end{aligned} \quad (10)$$

In Equation (10), $N = 2^{A-B}$, and $p(l)$ is the distribution function of random variable d , and:

$$p(l) = P\{d = l\} \quad (11)$$

Equation (10) can be reckoned as the cross-correlation function of signal $z(m)$ and distribution function $p(l)$. The frequency-domain expression can be written as:

$$Z'(e^{j\omega}) = Z(e^{-j\omega})P(e^{j\omega}) \quad (12)$$

In the following analysis, dither with Gaussian distribution and uniform distribution are introduced to demonstrate the restraint on truncation error.

4.1. Gaussian Dither Restrains Truncation Error

When dither is a discrete signal with Gaussian distribution [19], we need to work out the distribution function of Equation (11), i.e. $p(l)$.

Assume the continuous random variable $d(t)$ obeys Gaussian distribution, with average value of μ and variance of σ^2 , i.e. $d(t) \square N(\mu, \sigma^2)$. Suppose $d(t)$ and d abide by the relationship as:

$$d = \lfloor d(t) \rfloor \quad (13)$$

$\lfloor d(t) \rfloor$ means the biggest integer not greater than $d(t)$, then $p(l)$ can be achieved from the following equation:

$$p(l) = \int_l^{l+1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \quad (14)$$

The Fourier transformation of $p(l)$ is $P(e^{j\omega})$, and:

$$\begin{aligned} P(e^{j\omega}) &= \sum_{l=0}^{N-1} p(l)e^{-j\omega l} \\ &= \sum_{l=0}^{N-1} \int_l^{l+1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \square e^{-j\omega l} \end{aligned} \quad (15)$$

The amplitude-frequency characteristic is:

$$\begin{aligned}
|P(e^{j\omega})| &= \left| \sum_{l=0}^{N-1} \int_l^{l+1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt e^{-j\omega l} \right| \\
&\leq \sum_{l=0}^{N-1} \int_l^{l+1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt |e^{-j\omega l}| \\
&= \sum_{l=0}^{N-1} \int_l^{l+1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \\
&= \int_0^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt
\end{aligned} \tag{16}$$

Let $Q = \int_0^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt < 1$, where value of Q indicates the area among the range of $[0, N]$ below the probability density plot of $d(t)$. Value of Q will diminish, as variance σ^2 increases. Because $|P(e^{j\omega})| \leq Q < 1$, $|P(e^{j\omega})|$ will diminish with the increase of σ^2 . Thus, what can be inferred from Equation (12) is:

$$\begin{aligned}
|Z'(e^{j\omega})| &= |Z(e^{-j\omega})| |P(e^{j\omega})| \\
&\leq |Z(e^{-j\omega})| Q \\
&< |Z(e^{-j\omega})| = |Z(e^{j\omega})|
\end{aligned} \tag{17}$$

Equation (17) tells us that, because of the influence of Gaussian dither, the harmonic amplitude in truncation error spectrum is smaller than that without Gaussian dither. Meanwhile, the harmonic amplitude will decrease significantly with the increase of σ^2 , which efficiently illustrates dither's restraint on truncation error.

4.2. Uniform Dither Restrains Truncation Error

When the additive dither is a uniform distributed signal and $d \in [0, N-1] \cap Z$, the distribution function $p(l)$ can be achieved according to the definition of uniform distribution, [20].

$$p(l) = P\{d = l\} = \frac{1}{N} \tag{18}$$

Execute Fourier transformation towards Equation (18):

$$\begin{aligned}
P(e^{j\omega}) &= \sum_{l=0}^{N-1} p(l) e^{-j\omega l} \\
&= \sum_{l=0}^{N-1} \frac{1}{N} e^{-j\omega l} \\
&= \frac{1 - e^{-j\omega N}}{N(1 - e^{-j\omega})}
\end{aligned} \tag{19}$$

Then, the amplitude-frequency characteristic of the truncation error is:

$$\begin{aligned}
|Z'(e^{j\omega})| &= |Z(e^{-j\omega})| |P(e^{j\omega})| \\
&= \sum_k \frac{N}{k} \delta\left(-\omega - \frac{2\pi k}{N}\right) \left| \frac{1 - e^{-j\omega N}}{N(1 - e^{-j\omega})} \right| \\
&= \left| \sum_k \frac{1 - e^{-j\omega N}}{k(1 - e^{-j\omega})} \delta\left(-\omega - \frac{2\pi k}{N}\right) \right| \quad (20) \\
&= \left| \sum_k \frac{1 - e^{-j\omega - 2\pi k/N}}{k(1 - e^{-j\omega - 2\pi k/N})} \delta\left(-\omega - \frac{2\pi k}{N}\right) \right| \\
&= \left| \sum_k \frac{1 - e^{j\omega - 2\pi k/N}}{k(1 - e^{j\omega - 2\pi k/N})} \delta\left(-\omega - \frac{2\pi k}{N}\right) \right| \\
&= 0
\end{aligned}$$

In actual cases, because of the finite length effect of uniform distributed dither, the occurrence probability of each point in Equation (18) turns not out to be always the same and exists a certain difference from the ideal value of $1/N$. As a result, the actual value of $|Z'(e^{j\omega})|$ will not equal to zero, but only be close to zero. Anyhow, what can be inferred from the above analysis is that uniform distributed dither also has the ability to restrain truncation error effectively.

5. Simulation and Verification

Assume $x(n)$ is a sine signal with bit-width of $A = 14$ bits. The sample-rate is 1MSPS and frequency stands at 170kHz. Width of signal $y(n)$ obtained after truncation is $B = 10$. Then the width of truncation error $z(m)$ is $A - B = 4$.

Firstly, add a Gaussian dither with width of 4 bits into signal $x(n)$. Comparison chart in time-domain and frequency-domain of the truncation error $z(m)$ of signal without dither and $z'(m)$ of signal with dither can be shown in Figure 3 and Figure 4.

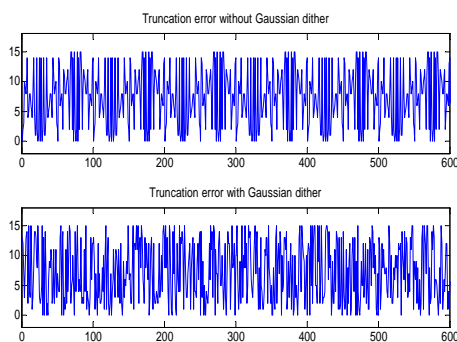


Figure 3. The Time-domain Comparison of Truncation Error without and with Gaussian Dither

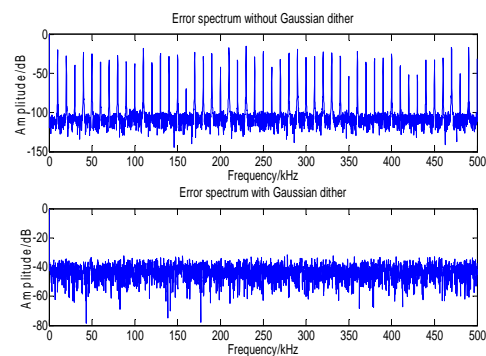


Figure 4. The Frequency-domain Comparison of Truncation Error without and with Gaussian Dither

Figure 3 tells the fact that waveform of truncation error without Gaussian dither in time-domain appears obvious periodicity, which leads to the harmonic distortion in the spectrum in Figure 4. The influence caused by harmonic distortion of truncation error is that, it will introduce a redundant peak signal in the spectrum of truncated signal $y(n)$. Performance of SFDR goes

worse owing to the mentioned influence. With additive dither, the periodicity is damaged, which helps decrease the harmonic distortion and increase SFDR of signal $y(n)$, as shown in Figure 5.

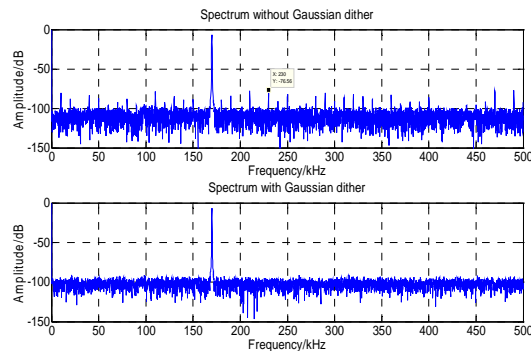


Figure 5. The Frequency-domain Comparison of Truncated Signal without and with Gaussian Dither

It can be inferred from Figure 5 that, SFDR of truncated signal with additive Gaussian dither increases about 16dB.

Put another two signal with frequency of 110kHz and 270kHz in $x(n)$, and add a uniform dither with 4-bits, then, the time-domain and frequency-domain comparisons of $z(m)$ and $z'(m)$ are described in Figure 6 and Figure 7 respectively.

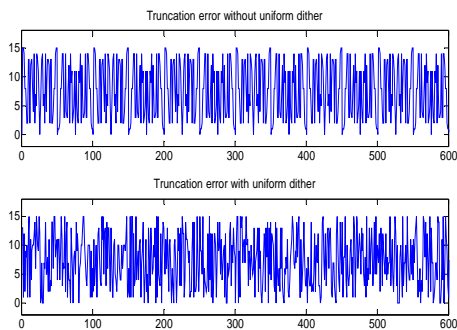


Figure 6. The time-domain Comparison of Truncation Error without and with Uniform Dither

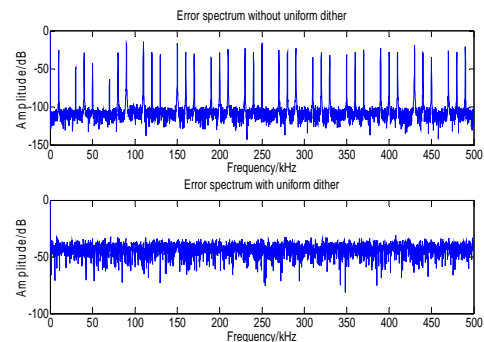


Figure 7. The Frequency-domain Comparison of Truncation Error without and with Uniform Dither

Same as what demonstrates in Figure 3, Figure 6 shows that the additive uniform dither damages signal's periodicity, which helps increase SFDR of truncated signal $y(n)$, as shown in Figure 8.

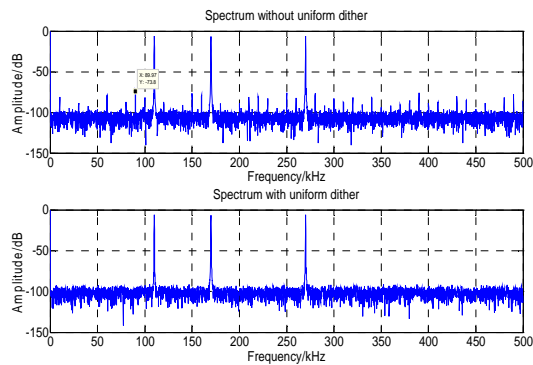


Figure 8. The Frequency-domain Comparison of Truncated Signal without and with Uniform Dither

It can be inferred from Figure 8 that, SFDR of truncated signal with additive uniform Dither increases about 15dB.

6. Conclusion

In this paper, we have presented a new way to analyze the truncation error in digital systems, especially in FPGA that referred above. Comparing to the analog dither technology in ADC-optimizing field, digital dither is imported into the analyzing process of truncation error, which helps restrain it and improve the system performance.

The proposed approach theoretically analyzes the peak noise introduced by truncation harmonic distortion, and deduces the mathematic expression of truncation error with additive dither based on statistical approach. In addition, MATLAB simulation of situations with additive Gaussian dither and uniform dither is presented. The simulation result reveals that, signals' SFDR performance will be increased significantly if we add another dither signal before operation of signals' precision conversion. The theoretically deduced expression keeps identical with simulation result.

References

- [1] Sanghamitra Roy, Prith Banerjee. An Algorithm for Trading off Quantization Error with Hardware Resources for MATLAB-Based FPGA Design. *IEEE Transactions on Computers*. 2005; 54(7): 886-895.
- [2] C Shi. Statistical Method for Floating-Point to Fixed Point Conversion. MS Thesis, Electrical Eng. And Computer Science Dept., Univ. of California, Berkeley. 2002.
- [3] KH Chang, WG Bliss. Finite Word-Length Effects of Pipelined Recursive Digital Filters. *IEEE Trans. Signal Processing*. 1994; 42(8): 1983-1995.
- [4] LB Jackson, KH Chang, WG Bliss. Comments on 'Finite Word-Length Effects of Pipelined Recursive Digital Filters. *IEEE Trans. Signal Processing*. 1995; 43(12): 3030-3032.
- [5] RM Gray, DL Neuhoff. Quantization. *IEEE Trans. Information Theory*. 1998; 44(6): 2325-2383.
- [6] W Sung, KI Kum. Simulation-Based Word-Length Optimization Method for Fixed-Point Digital Signal Processing Systems. *IEEE Trans. Signal Processing*. 1995; 43(12): 2209-2212.
- [7] S Kim, WSung. Fixed-Point Error Analysis and Wordlength Optimization of a Distributed Arithmetic Based 8x8 2D-IDCT Architecture. Proc. Workshop VLSI Signal Processing. 1996: 398-407.
- [8] H Keding, M Willems, M Coors, H Meyr. FRIDGE: A Fixed-Point Design and Simulation Environment. Proc. Design, Automation, and Test in Europe. 1998: 429-435.
- [9] Dias P, Silva G, Cruz S. Dithering performance of over sampled ADC systems affected by hysteresis. Journal of the International Measurement Confederation. 2002; 32(1): 51-59.
- [10] Wagdy Z, Fawzy M. Effect of additive dither on the resolution of ADC's with single-bit or mulibit errors. *IEEE Transactions on Instrumentation and Measurement*. 1996; 45(2): 610-615.
- [11] Suresh B, Wollman HB. Testing an ADC linearized with pseudorandom dither. *IEEE Transactions on Instrumentation and Measurement*. 1998; 47(4): 839-848.
- [12] Zhang Yun, Li Guangjun. A pipelined ADC structure adaptable to dither introduction. Modern Electronics Technique. 2011; 34(10): 160-162.

- [13] Blesser B, Locantii B. *The application of narrowband Dither operating at the Nyquist frequency in digital systems to provide improved signal to noise ratio over conventional Dithering*. *Audio Eng.* 1987; 35(6): 446-454.
- [14] Anna D. A-D conversion with Dither signal-possibilities and limitations. *Measurement Science Review*. 2001; 1(1): 75-78.
- [15] Shu YS, Song BS. A 15 bit linear 20MSample/spipelined ADC digitally calibrated with signal-dependent Dithering. *IEEE Journal Solid-State Circuits*. 2008; 43(2): 342-350.
- [16] Yu CH H, LI JL. A White Noise Filtering Method for DOA Estimation of Conherent Signals under Low SNR. *Signal Processing*. 2012; 28(7): 957-962.
- [17] Wang LB, Cui CH, Sha ZH H. Sparse Decomposition Method of Smooth Signal Under Truncation Effect. *Signal Processing*. 2011; 27(6): 956-960.
- [18] Chen TQ, Xu J, Zhu K. Error Analysis and System Design of High-Accuracy Pipelined A/D Converters. *Microelectronics*. 2008; 38(1): 126-128.
- [19] Cheng M ZH, Jing WP. Design and Analysis of a Novel Pipelined ADC. *Journal of University of Electronic Science and Technology of China*. 2008; 37(6): 930-933.
- [20] Wagdy MF, Ng W. Validity of uniform quantization error model of sinusoidal signals without and with Dither. *IEEE Transactions on Instrumentation and Measurement*. 1989; 38(3): 718-722.