Rigorous Analysis of Quantization Error of an A/D Converter Based on β-Map

Takaki Makino*, Yukiko Iwata†, Yutaka Jitsumatsu‡, Masao Hotta§, Hao San§ and Kazuyuki Aihara*
*Institute of Industrial Science, The University of Tokyo, Tokyo, 153–8505 Japan Email: mak@sat.t.u-tokyo.ac.jp
†FIRST, Aihara Innovative Mathematical Modelling Project, Japan Science and Technology Agency
‡Kyushu University, Fukuoka, 819–0395 Japan §Tokyo City University, Tokyo, 158–8557 Japan

***Abstract*—A non-binary analog-to-digital converter (ADC) based on β-expansion, called β-encoder, is so far reported to achieve robustness against large process variation and wide environment change. Quantization error of β-encoder is not uniformly distributed, which makes the mean squared error (MSE) evaluation difficult. An analysis method for giving the upper bound of MSE of the quantization error is provided. Signal-to-noise-ratio (SNR) is also evaluated and the result is effective for designing β-encoders*.**

***Index Terms*—Analog-to-Digital and Digital-to-Analog converters, quantization error, β-map, mean squared error.**

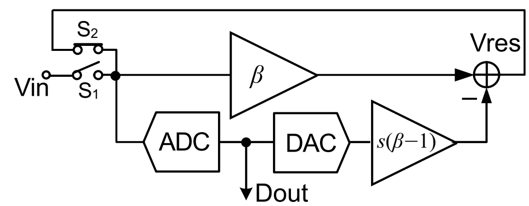

Fig. 1: Simplified block diagram of the β-encoder (cyclic ADC based on β-expansion). In this paper we use the ordinary β-map, which means that the full-scale width s is set to $\frac{1}{\beta-1}$.

I. Introduction

Robust analog-to-digital converter (ADC) architecture is desirable for not only nowadays mixed signal LSIs but also next generation CMOS technology, to meet the requirement of ADC performances of high sampling frequency, high resolution, small chip area (the same word as low cost), and low power consumption. In the conventional binary architecture, the linearity of ADC is very sensitive to the accuracy of analog components. Therefore, high-gain wideband amplifiers as well as high accuracy matched devices, such as transistors, capacitors, and resistors, are necessary to satisfy the required ADC linearity, leading to large chip area and high power consumption.

A recently proposed architecture, called β-encoder [1], is a non-binary ADC based on β-expansion, which has a self-correction property for fluctuations of amplifier factor β and quantizer threshold ν. It is shown [2] that the circuit based on the approach (Fig. 1) has robustness that tolerates the conversion errors caused by finite gain of amplifier and mismatches of the devices, and that the proposed β-estimation algorithm eliminates the need for any digital calibration technique. Just by adding a simple conversion sequence with the effective radix-value β, we can realize a reliability-enhanced ADC with greatly relaxed power and area penalties for high-gain amplifier and high-accuracy circuit elements.

From a viewpoint of circuit design, it is important to give a theoretical evaluation of mean squared error (MSE) and signal-to-noise ratio (SNR), which guarantee the accuracy and linearity of β-encoders. The only known theoretical results about the quality of β-encoder is its maximum quantization error [1], [3]. The purpose of this paper is to make a theoretical derivation of simple and accurate MSE and SNR evaluation.

The fact that the β-expansion map $C_{\beta,\nu}(x)$ is locally eventually onto $[\nu-1,\nu]$ implies that we may evaluate MSE by assuming that $C^L_{\beta,\nu}(X)$ is uniformly distributed in the interval, where X is a random variable for x. Such an assumption makes the MSE to be $\frac{1}{12}\beta^{-2L}$, where L is the number of bits[1]. Computer simulation, however, shows that the numerically calculated MSE is deviated from such an evaluation and is not sufficient to guarantee the quality of a β-encoder.

The trajectory of $C^i_{\beta,\nu}(x)$ for $i = 0, 1, \ldots$ with $x \in [\nu-1, \nu]$ follows the Parry's invariant density [4]. Hence, it seems a good way to evaluate MSE based on the Parry's density. However, this has two difficulties. One is that the main target of L is from 12 to 16 in real applications. Such a size of L is not sufficiently large for the density function of quantization error to converge to the invariant measure. The other is that Parry's invariant density is expressed as an infinite sum of $\pm\beta^{-n}$s, which complicates the MSE evaluation based on the direct application of Parry's density. This situation motivated us to develop a new MSE analysis method.

We provide a method for analyzing the MSE of β-encoder by introducing a notion of *segments*, i.e., linear pieces within the L-nested β-map. Such a method, called level j truncation, enables us to give a tight upper bound of MSE. Using this upper bound, and restricting our attention to the cautious map with $\nu = \frac{\beta}{2(\beta-1)}$, we prove that the MSE of such β-encoder is smaller than $\frac{1}{12}\beta^{-2L}$ if $\frac{1+\sqrt{5}}{2} \le \beta \le 2$ and $5 \le L \le 18$. Such a sufficient condition is effective for designing β-encoders. We

*This research is supported by the Aihara Innovative Mathematical Modelling Project, the Japan Society for the Promotion of Science (JSPS) through the "Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program)", initiated by the Council for Science and Technology Policy (CSTP).

[1] In case of scale-adjusted β-map with a full-scale s, MSE is multiplied by $s^2(\beta-1)^2$. For ordinary β-map, $s = \frac{1}{\beta-1}$.

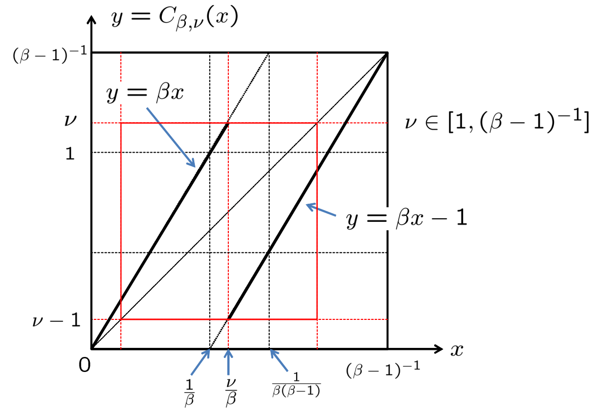

Fig. 2: Ordinary β-map $C_{\beta,\nu}(x)$.

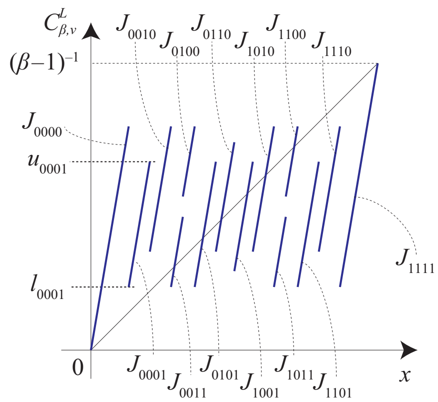

Fig. 3: Segments in the map $C^L_{\beta,\nu}$. With $L = 4$, there are 14 segments indexed by $\mathbf{b} \in \{0000, \ldots, 1111\}$. Note that segments J_{0111} and J_{1110} have zero length and were not drawn. The lower and upper ends of each segment are given as $l_{\mathbf{b}}$ and $u_{\mathbf{b}}$, respectively.

also provide evaluation of SNR as well.

II. Preliminaries

A β-map $C_{\beta,\nu} : \left[0, \frac{1}{\beta-1}\right] \to \left[0, \frac{1}{\beta-1}\right]$, illustrated in Fig. 2, is defined by

$$C_{\beta,\nu}(x) = \begin{cases} \beta x & \beta x < \nu, \\ \beta x - 1 & \beta x \geq \nu, \end{cases} \tag{1}$$

where $1 < \beta \leq 2$ is an amplification factor, and $1 \leq \nu \leq \frac{1}{\beta-1}$ is a threshold. Its associated bit sequence is defined as $b_{i,C_{\beta,\nu}} = 0$ if $C^{i-1}_{\beta,\nu}(x) < \beta^{-1}\nu$ and $b_{i,C_{\beta,\nu}} = 1$ if $C^{i-1}_{\beta,\nu}(x) \geq \beta^{-1}\nu$. A decoder reconstructs a signal x from the bit sequence. The decoded value is equal to the midpoint of the subinterval [3], i.e., $\hat{x}^L = \sum_{i=1}^{L} b_{i,C_{\beta,\nu}}\beta^{-i} + \left(\nu - \frac{1}{2}\right)\beta^{-L}$. Then the quantization error is expressed as

$$|x - \hat{x}^L| = \beta^{-L}\left|C^L_{\beta,\nu}(x) - \left(\nu - \frac{1}{2}\right)\right|. \tag{2}$$

The map is called greedy-, lazy-, and cautious-expansion, respectively, if $\nu = 1$, $\nu = \frac{1}{\beta-1}$, and $1 < \nu < \frac{1}{\beta-1}$.

The β-encoder is robust against a fluctuation of the threshold under the assumption $\nu \in [1, \frac{1}{\beta-1}]$ and it can work for unknown ν properly [1]. In order to analyze the MSE of a β-encoder, however, we suppose ν is fixed to a known value, as well as we have to assume some distribution function for the input signal x. Uniform distribution on the full-scale, $[0, \frac{1}{\beta-1}]$ is assumed in this paper.

Then we define the MSE as follows:

$$\begin{aligned} \mathrm{MSE}(C_{\beta,\nu}, L) &= \int_0^{\frac{1}{\beta-1}} |x - \hat{x}^L|^2 \, \frac{1}{\frac{1}{\beta-1}} dx \\ &= (\beta-1)\beta^{-2L}\int_0^{\frac{1}{\beta-1}} \left(C^L_{\beta,\nu}(x) - \left(\nu - \frac{1}{2}\right)\right)^2 dx \end{aligned} \tag{3}$$

We assume the input is uniformly distributed on the full-scale $[0, \frac{1}{\beta-1}]$, but the following analysis can be applied to other distributions with a slight modification.

III. Analysis of MSE

In this paper, we propose a new method of analyzing $C^L_{\beta,\nu}$ based on *segments* (Fig. 3). $C^L_{\beta,\nu}$ consists of at most 2^L linear segments, indexed by bit sequence $\mathbf{b} = b_1 \cdots b_L$. We denote the location of each segment of $C^L_{\beta,\nu}$ by $J_{\mathbf{b}} = [l_{\mathbf{b}}, u_{\mathbf{b}}]$, where $l_{\mathbf{b}}$ and $u_{\mathbf{b}}$ are the lower and upper ends of the projection of the segment on y-axis, respectively. Segments can be used to calculate MSE by changing variable in Eq. (3):

$$\begin{aligned} &\mathrm{MSE}(C_{\beta,\nu}, L) \\ &\quad= (\beta-1)\beta^{-2L} \sum_{\mathbf{b}\in\{0,1\}^L} \int_{l_{\mathbf{b}}}^{u_{\mathbf{b}}} \left(y - \left(\nu - \frac{1}{2}\right)\right)^2 \frac{dy}{\beta^L} \\ &\quad= (\beta-1)\beta^{-3L} \sum_{\mathbf{b}\in\{0,1\}^L} f(J_{\mathbf{b}}), \end{aligned} \tag{4}$$

where $f(J_{\mathbf{b}}) = \frac{1}{3}(u_{\mathbf{b}} - (\nu - \frac{1}{2}))^3 - \frac{1}{3}(l_{\mathbf{b}} - (\nu - \frac{1}{2}))^3$.

Equation (4) implies that $\sum_{\mathbf{b}\in\{0,1\}^L} \beta^{-L}\mathbb{1}_{[l_{\mathbf{b}}, u_{\mathbf{b}}]}$ is considered as the Lth iteration of the Perron-Frobenius operator corresponding to $C_{\beta,\nu}$ with respect to the initial distribution $(\beta-1)\mathbb{1}_{[0,\frac{1}{\beta-1}]}$, where $\mathbb{1}_{[a,b]}(x)$ is 1 if $a \leq x \leq b$ and 0 otherwise. Such a distribution function converges to $h(x - \nu + 1)$ as $L \to \infty$, where $h(x)$ is the Parry's invariant density [4] under the (β, α)-transformation $T_{\beta,\alpha}$: $[0, 1) \mapsto [0, 1)$ defined by $T_{\beta,\alpha} = \beta x + \alpha \bmod 1$, $\beta \geq 1$ and $0 \leq \alpha \leq 1$.

Figure 3 shows an example of fourth iterated cautious β map (i.e., $\nu = \frac{\beta}{2(\beta-1)}$). It should be noted that two segments indexed by 0111 and 1000 do not appear. This phenomenon occurs if segments labelled by 011 and 100 in $L = 3$ are included in $[0, \frac{\nu}{\beta}]$ and $[\frac{\nu}{\beta}, \frac{1}{\beta-1}]$, respectively. The condition whether a segment disappears or not depends on the value of β and L. Such a situation makes the analysis of MSE difficult. We, however, provide a method to perform such kind of calculation effectively by analyzing the pattern of $l_{\mathbf{b}}$ and $u_{\mathbf{b}}$ rigorously, as shown in the next section. The analysis shows that the summation in Eq. (4) with at most 2^L terms can be reduced to a summation with only $2L$ terms.

A. Exact MSE for the uniform distribution

We denote a vector of segments in $C^i_{\beta,\nu}$ by

$$\mathrm{SEG}(i) = \langle \overbrace{J_{\underbrace{0\cdots0}_{i}}, \ldots, J_{1\cdots1}}^{2^i} \rangle. \tag{5}$$

Let the initial segment ($i = 0$) be the full-scale of the β-map, i.e., $\mathrm{SEG}(0) = \langle [0, \frac{1}{\beta-1}] \rangle$. For $i = 1$, We can derive. $\mathrm{SEG}(1) = \langle J_0, J_1 \rangle = \langle [0, \nu], [\nu - 1, \frac{1}{\beta-1}] \rangle$. For general $i \geq 2$, the exact set of segments are calculated by iteration:

Proposition 3.1: We can compute $\mathrm{SEG}(i{+}1)$ from $\mathrm{SEG}(i)$:

$$l_{\mathbf{b}0} = \min(\nu, \beta l_{\mathbf{b}}) \ , \qquad l_{\mathbf{b}1} = \max(\nu - 1, \beta l_{\mathbf{b}} - 1) \ ,$$
$$u_{\mathbf{b}0} = \min(\nu, \beta u_{\mathbf{b}}) \ , \qquad u_{\mathbf{b}1} = \max(\nu - 1, \beta u_{\mathbf{b}} - 1) \ .$$

Note that, the length of a segment $|J_{\mathbf{b}}| = u_{\mathbf{b}} - l_{\mathbf{b}}$ may become zero and disappear from the graph of $C^L_{\beta,\nu}(x)$. Such a segment has no effect on the evaluation of MSE, because if $|J_{\mathbf{b}}| = 0$, then $f(J_{\mathbf{b}}) = 0$ and $|J_{\mathbf{b}0}| = |J_{\mathbf{b}1}| = 0$.

The vector may contain several non-zero-length segments that have the identical location. Indeed, we can show the following property:

Lemma 3.2: Let us denote the number of distinct, non-zero-length segment locations in $\mathrm{SEG}(i)$ as $\#\mathrm{SEG}(i)$. Then

$$\#\mathrm{SEG}(i+1) \leq \#\mathrm{SEG}(i) + 2. \tag{6}$$

Consequently, $\#\mathrm{SEG}(1) = 2$ gives $\#\mathrm{SEG}(i) \leq 2i$.
The proof is omitted for lack of space.

Thus we can give a numerical index for the possible locations of segments, $J_k = [l_k, u_k]$ $(k \geq 0)$, such that $J_{\mathbf{b}} = J_{k_{\mathbf{b}}}$, where $k_{\mathbf{b}} \in \{0, \ldots, 2i-1\}$ for $\mathbf{b} \in \{0,1\}^i$. This means that, to calculate the MSE for Lth iteration, we do not have to compute locations of 2^L segments, but only need counting the number $n_k^{(L)}$ of segments on the kth location J_k for $k = 0, \ldots, 2L-1$. The counting can be done from some recurrence formula with respect to L by use of the property of $n_k^{(L)}$. Thus we can derive MSE from Eq. (4):

$$\mathrm{MSE}(C_{\beta,\nu}, L) = (\beta - 1)\beta^{-3L} \sum_{k=0}^{2L-1} n_k^{(L)} f(J_k). \tag{7}$$

B. Truncating enumeration of distinct segments

Eq. (7) gives the exact MSE. The number of segments $n_k^{(L)}$, however, changes sensitively by the values of β and ν. This fact prevents us to analyze the MSE, because fluctuation of β is allowed in β-encoders. We restrict the number of distinct segments up to $2j$, where $j < L$. Such a method, called level j truncation, helps us to analyse the MSE for a range of β and ν.

We considered the segments that escaped from the considered locations as "lost." Specifically, the set of the indices of lost segments at ith iteration at the level j truncation is as follows: For $\mathbf{b} = b_1 \cdots b_i \in \{0,1\}^i$,

$$\mathrm{LOST}(i,j) = \{\mathbf{b} \in \{0,1\}^i : \exists i' \leq i \text{ s.t. } k_{b_1 \cdots b_{i'}} \geq 2j\}. \tag{8}$$

Thus we consider the number of non-lost segments for $k = 0, \ldots, 2j-1$ as well as the number of lost segments:

$$n_k^{(i,j)} = \#\{\mathbf{b} | \mathbf{b} \notin \mathrm{LOST}(i,j), J_{\mathbf{b}} = J_k\}, \tag{9}$$
$$n_{lost}^{(i,j)} = \#\mathrm{LOST}(i,j). \tag{10}$$

This truncation method has the following advantages:

- Given β, ν and j, we can derive some recurrence formula for $n_k^{(i,j)}$ and $n_{lost}^{(i,j)}$ with respect to i considering only $2j+1$ variables.
- Approximated $n_k^{(i,j)}$ values are constant for a range of β and ν.

Let h_{lost} be the average of the length of segments belonging to the set of lost segments, $\mathrm{LOST}(L,j)$, defined by

$$h_{lost} = \frac{\sum_{\mathbf{b} \in \mathrm{LOST}(L,j)} (|J_{\mathbf{b}}|)}{n_{lost}^{(L,j)}} = \frac{\frac{\beta^L}{\beta-1} - \sum_{k=0}^{2j-1} n_k^{(L,j)} |J_k|}{n_{lost}^{(L,j)}}$$

for $n_{lost}^{(L,j)} \neq 0$. Jensen's inequality gives

$$\sum_{\mathbf{b} \in \mathrm{LOST}(L,j)} f(J_{\mathbf{b}}) \leq n_{lost}^{(L,j)} g\left(h_{lost}\right), \tag{11}$$

where

$$g(x) = \begin{cases} \frac{1}{16}x + \frac{1}{48} & (x \geq \frac{1}{4}), \\ \frac{1}{3}\left(x - \frac{1}{2}\right)^3 + \frac{1}{24} & (x < \frac{1}{4}), \end{cases}$$

is a convex function and $f(J_{\mathbf{b}}) \leq g(|J_{\mathbf{b}}|)$ is satisfied for all $l_{\mathbf{b}} \geq \nu - 1$ and $u_{\mathbf{b}} \leq \nu$. Then, we give an upper bound using the jth level truncation as follows:

$$\mathrm{MSE}\left(C_{\beta,\nu}, L\right) \leq (\beta - 1)\beta^{-3L} \left(\sum_{k=0}^{2j-1} n_k^{(L,j)} f(J_k) + n_{lost}^{(L,j)} g(h_{lost}) \right). \tag{12}$$

This bound becomes tighter when j goes closer to L. When $j = L$, $n_{lost}^{(L,j)}$ becomes 0 and the bound becomes equivalent to Eq. (7).

C. A simple upper bound for MSE

Hereafter, we assume a *cautious* β-expansion with $\nu = \frac{\beta}{2(\beta-1)}$. The MSE is deviated from the intuitive expression $\frac{1}{12}\beta^{-2L}$ and can be greater than this value for some β and L. Using Inequality (12) with level 8 truncation, we give a sufficient condition for the MSE to be less than this value, as follows:

Theorem 3.3: For $\frac{1+\sqrt{5}}{2} \leq \beta \leq 2$ and $5 \leq L \leq 18$,

$$\mathrm{MSE}\left(C_{\beta, \frac{\beta}{2(\beta-1)}}, L\right) \leq \frac{1}{12}\beta^{-2L}. \tag{13}$$

Note that the ranges of β and L are just a sufficient condition for Inequality (13). Simulation results suggest that the same inequality holds for other cases, such as $L > 18$.

IV. SNR OF β-ENCODERS

We have analyzed the MSE of β-encoders. However, signal-to-noise ratio (SNR) is used as a criterion for the quality of ADCs more often than MSE. We can roughly evaluate the SNR from the MSE.

Suppose a sinusoidal waveform of a certain frequency f_{in} is input to the ADC. Such a signal is sampled and quantized. SNR is defined, in a frequency domain, as a ratio of the energy of the signal component at frequency f_{in} to the total energy of quantization noise components. The range of β-encoder is $[0, \frac{1}{\beta-1}]$. Then, the energy of the input signal for one period is $(\beta-1)^{-2}/8$. The distribution of the sampled sinusoidal waves

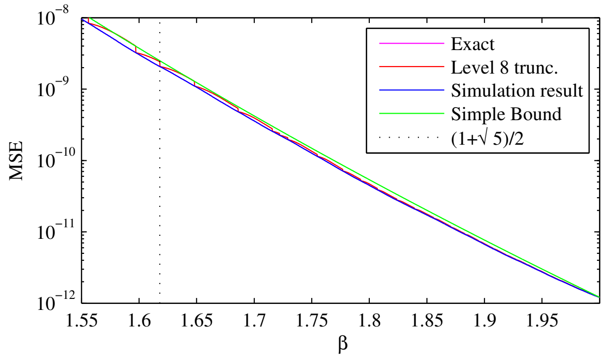

Fig. 4: Comparison of MSE with simulation results, $L = 18$.

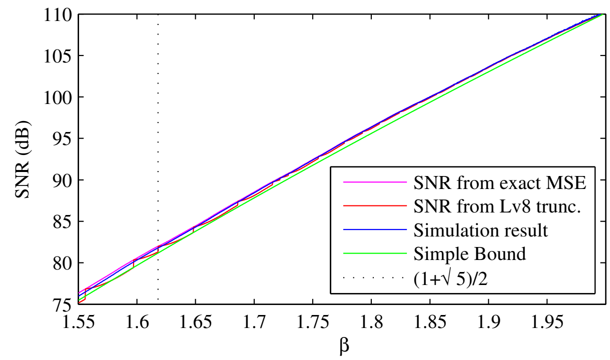

Fig. 6: Comparison of SNR (dB) with simulation results, $L = 18$.

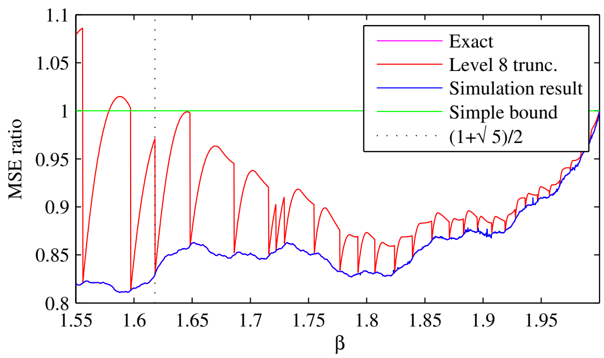

Fig. 5: Comparison of MSE ratio (with simple bound = 1), $L = 18$.

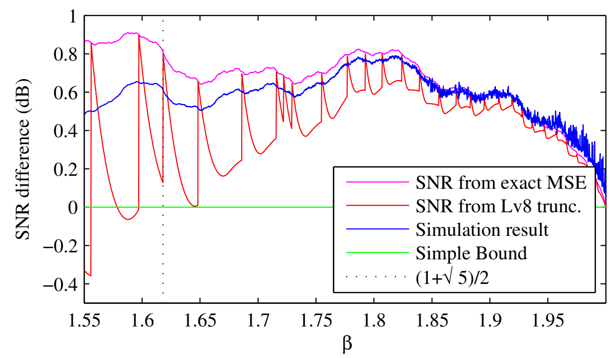

Fig. 7: Comparison of SNR difference (simple bound = 0), $L = 18$.

is not uniform, but we approximate it as uniform. Then the noise component N is approximated by the MSE. We obtain

$$\mathrm{SNR} = 10\log_{10}\frac{S}{N} \approx 10\log_{10}\frac{(\beta-1)^{-2}/8}{\mathrm{MSE}(C_{\beta,\nu},L)}(\mathrm{dB}). \quad (14)$$

Using Eq. (13), we get an approximated lower bound of SNR:

$$\mathrm{SNR} \gtrapprox 20L\log_{10}\beta - 20\log_{10}(\beta-1) + 1.76\ (\mathrm{dB}). \quad (15)$$

For a scale-adjusted β map with its scale s, both signal and noise components are multiplied by $s^2(\beta-1)^2$. Thus, Eqs. (14) and (15) holds for any s.

V. Simulation Results

Figures 4 and 5 show the comparison of the derived MSE values and the results of numerical simulation. Since the simulation results match perfectly to the exact MSE value (Eq. (7)), they are not distinguishable in the plot. The MSE based on level 8 truncation is slightly worse than them, but for the range $\frac{1+\sqrt{5}}{2} \leq \beta < 2$, it is better than the simple bound (Eq. (13)).

Figures 6 and 7 show the comparison of the SNR values with simulation using sinusoidal wave input. The SNR calculated from the simulation is slightly worse than SNR calculated from the exact MSE and Eq. (14), but it is better than the simple bound (Eq. (15)) for $\frac{1+\sqrt{5}}{2} < \beta < 2$. Moreover, the difference between the simulation and the simple bound is less than 0.8 dB. This means that Eq. (15) can be used as a good approximation of the ideal SNR value of the β-encoder.

VI. Concluding Remarks

A method providing a rigorous analysis of the MSE of ADCs based on β-map was presented. Using this method, we gave a sufficient condition for MSE to be less than $\frac{1}{12}\beta^{-2L}$. Such an MSE analysis guarantees the quality of β-encoders and leads to a useful SNR evaluation. Uniform distribution on the full-scale $[0, \frac{1}{\beta-1}]$ for the input value x has been assumed in this paper. However, the proposed analysis method can be applied to other distribution functions, too. This topic as well as the proofs of Lemma 3.2 and Theorem 3.3 will be given in a separate paper.

Acknowledgement

The authors would like to thank Professor T. Kohda, Professor Y. Horio, and Professor Y. Takahashi, for their very helpful comments and discussions.

References

[1] I. Daubechies, R.A. DeVore, C.S. Güntürk, and V. A. Vaishampayan, “A/D Conversion With Imperfect Quantizers,” *IEEE Trans. Inform. Theory*, vol. 52, no.3, pp. 874–885, Mar. 2006.

[2] T. Maruyama, H. San and M. Hotta, “Robust Switched-Capacitor ADC based on β-expansion,” *2011 IEEJ Int. Analog VLSI Workshop*, Bali, Indonesia, pp. 171–175, Nov. 2011.

[3] T. Kohda, Y. Horio, Y. Takahashi, and K. Aihara, ”Dynamics and Implementation of β encoders,” *Mathematical Engineering Technical Reports*, vol. METR 2012-07, pp. 1–67, 2012.

[4] W. Parry, ”Representations for real numbers,” *ActaMath. Acad. Sci. Hungar.*, vol. 15, pp. 95–105, 1964.