Rigorous estimates of quantization error for A/D converters based on beta-map

Takaki Makino, Yukiko Iwata, Katsutoshi Shinohara, Yutaka Jitsumatsu, Masao Hotta, Hao San, and Kazuyuki Aihara

1 Institute of Industrial Science, the University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
2 Meteorological College, Japan Meteorological Agency, 7-4-81 Asahicho, Kashiwa, Chiba 277-0852, Japan
3 FIRST Aihara Innovative Mathematical Modelling Project, Japan Science and Technology Agency
4 Kyushu University, Fukuoka 819-0395, Japan
5 Tokyo City University, Tokyo 158-8557, Japan

Received March 26, 2014; Revised June 16, 2014; Published January 1, 2015

Abstract: A β-encoder, a non-binary analog-to-digital converter that is based on the β-expansion, is reported to be robust to large process variations and widespread environment fluctuations. However, the quantization error of the β-encoder is complicatedly distributed and hard to be estimated due to the chaotic nature of the β-transformation. In this work, we propose an analysis method for determining an upper bound of the mean squared error of the quantization. We also provide an evaluation of the signal-to-quantization-noise ratio, which is useful for designing β-encoders.

Key Words: analog-to-digital converters, digital-to-analog converters, quantization error, β-map, mean squared error

1. Introduction

A new analog-to-digital converter (ADC) architecture called β-encoder [1, 2], is attracting attention due to its intrinsic robustness against fluctuations of analog components. The β-encoder, explained by the deterministic dynamics of β-map (also known as the “multi-valued Rényi-Parry map [3–5],” see [6]), has a self-correction property for fluctuations of the amplifier factor β and quantizer threshold.
Fig. 1. Simplified block diagram of the $\beta$-encoder (cyclic ADC based on the $\beta$-expansion). In this paper we consider the ordinary $\beta$-map, which means that the full-scale width $s$ is $\frac{1}{\beta - 1}$.

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

parameter $\nu$, without additional analog components such as comparators in overrange recovery architecture [7]. It has been shown in [8, 9] that circuits based on this approach (Fig. 1), together with a simple conversion sequence from an effective radix-value $\beta$, results in a reliability-enhanced ADC without depending on high-gain wideband amplifiers and high-accuracy circuit elements. Furthermore, the proposed $\beta$-estimation algorithm [9–11] drastically simplifies the digital calibration process.

For use in practical applications, however, we need a theoretical evaluation of the several important performance measures for designing ADCs, such as mean squared errors (MSEs) and signal-to-quantization-noize ratios (SQNRs). In particular, such a theoretical estimation is indispensable for designing a $\beta$-encoder circuit to meet specific demands, in which the beta value $\beta$ and the number of bits $L$ of the encoder must be decided so as to satisfy required MSE or SQNR bounds. However, because of the chaotic behavior of the underlying $\beta$-map $C_{\beta,\nu}$ (Fig. 2; see Section 2 for the precise definition), mathematical analysis of $\beta$-encoder is complicated compared to the traditional binary ADCs. Previously known results for the quality of the $\beta$-encoder refer to its maximum quantization error [2, 12], which is not enough to estimate the MSE or the SQNR of the encoder. In this paper, we consider a theoretical upper bound of MSE and an estimate of SQNR of an $L$-bit $\beta$-encoder under the following assumptions: (1) the $\beta$ value in the encoder is accurately known, (2) the encoder is working as designed (i.e., device mismatches are ignored), and (3) the $\beta$ value is fixed (i.e., the ADC is cyclic).

In our previous work [13], we proposed a new analysis method for the MSE of a $\beta$-encoder. Our analysis is based on the notion of segments, linear pieces in a nested $\beta$-map, which allows us to evaluate the integration over the density function of $L$-nested $\beta$-map as a sum of only $2L$ terms. Using this method, together with a technique called level-$j$ truncation, we claimed Theorem 4.1 (see Section 4.3 of this paper), that is, $\frac{1}{12\beta^{-2L}}$ gives an upper bound of the MSE if $1 + \sqrt{5} \leq \beta \leq 2$, $5 \leq L \leq 18$, and $\nu = \frac{\beta}{2(\beta - 1)}$. However, in the previous paper, we omitted some important parts of the proofs because of the page limit.

The main goal of this paper is to provide these missing proofs in our previous work. We give the rigorous proofs for Proposition 3.1 and Lemma 3.3, which are the foundation of our analysis. We also cover the exact recurrence formula for number of segments that can be used for truncated and non-truncated cases (Eqs. (33)–(35)). These formulas are required to calculate the exact values of the truncated MSE (Eq. (39)). Moreover, we supply the Appendix, which includes the proof of Theorem 4.1 in one specific case; repeating similar elementary analytic arguments for many sub-cases constitutes the complete proofs of Theorem 4.1. We also added the discussion of the parameter range based on new simulation results in Section 6.

The rest of the paper is organized as follows. We first give preliminaries in Section 2. Then we give the proofs of key propositions and lemmas (Section 3). This leads us to the analysis of MSE (Section 4) and SQNR (Section 5). Then we discuss the results of numerical simulations in Section 6.
2. Preliminaries

A $\beta$-map $C_{\beta, \nu} : [0, \frac{1}{\beta-1}) \to [0, \frac{1}{\beta-1})$, illustrated in Fig. 2, is defined by the following formula:

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

where $1 < \beta \leq 2$ is an amplification factor, and $\nu$ with $1 \leq \nu \leq \frac{1}{\beta-1}$ is a threshold. For each $x \in [0, \frac{1}{\beta-1})$, its associated bit sequence $b_{\beta, \nu}(x) = b_{\beta, \nu}(i)_{i \geq 1}$ is defined as $b_{i, \beta, \nu} = \mathbb{I}[C_{\beta, \nu}(x) \geq \beta^{-i}]$, where $\mathbb{I}[\cdot]$ is 1 when the condition is true and 0 otherwise.

Given a bit sequence $b$ of length $L$, a decoder produces an estimate of its original signal as follows:

$$\hat{x}_L(b) = \sum_{i=1}^{L} b_i \beta^{-i} + \theta \beta^{-L},$$

where the term $\theta \beta^{-L}$ is added to compensate the error caused by quantization. In this term, $\theta$ is a fixed constant which is independent of the bit sequence $b$. The quantization error can be reduced if we appropriately choose the parameter $\theta$, more precisely, to be the expected value of the residual after $L$-step encoding. In this article, we consider the decoding scheme for which $\theta$ is equal to the midpoint of the invariant subinterval $[\nu - 1, \nu]$. We denote the point by $\theta_0$, i.e.,

$$\theta_0 = \nu - \frac{1}{2}.$$  \hspace{1cm} (3)

The quantization error is expressed as

$$|x - \hat{x}_L| = \beta^{-L} |C_{\beta, \nu}(x) - \theta|. \hspace{1cm} (4)$$

The map is called greedy-, lazy-, or cautious-expansions for $\nu = 1$, $\nu = \frac{1}{\beta-1}$, and $1 < \nu < \frac{1}{\beta-1}$, respectively [15].

The $\beta$-encoder is robust against fluctuations of the threshold under the assumption that $\nu \in [1, \frac{1}{\beta-1}]$, and it also works for unknown $\nu$ [2]. However, to analyse the MSE of the quantization error we suppose that $\nu$ is fixed at a known value. We also assume that the signal $x$ obeys some distribution law, that is, the input is uniformly distributed on the full-scale $[0, \frac{1}{\beta-1})$. However, the following analysis can be also applied to other distributions as well with a slight modification.

Then, the MSE is given as follows:

$$\text{MSE}(C_{\beta, \nu}, L, \theta) = \int_{0}^{\frac{1}{\beta-1}} |x - \hat{x}_L|^2 \frac{1}{\beta-1} \, dx$$

$$= (\beta - 1) \beta^{-2L} \int_{0}^{\frac{1}{\beta-1}} \left(C_{\beta, \nu}(x) - \theta \right)^2 \, dx.$$  \hspace{1cm} (5)

This means that the problem of calculating the MSE is reduced to the evaluation of an integral containing $C_{\beta, \nu}^L$ (an $L$-nested $\beta$-map). The integral is, however, hard to be evaluated directly, since $C_{\beta, \nu}^L$ contains up to $2^L$ discontinuities. The rest of the paper pursues an appropriate way to approximate $C_{\beta, \nu}^L$ that gives a strict upper bound to the MSE.

One of the easiest approach is to use the fact that the $\beta$-map $C_{\beta, \nu}$ has an invariant interval $[\nu - 1, \nu]$ as an attractor. Under a simplified assumption that $C_{\beta, \nu}(X)$ is uniformly distributed in the invariant interval, we can obtain an approximated evaluation of the MSE as $\frac{1}{L} \beta^{-2L}$, where $L$ is the number of bits. It is, however, impossible to guarantee the quality of a $\beta$-encoder based on this approach. Indeed we observe that the evaluation has non-negligible gap from the numerical calculation by computer simulation.

Another possible approach is to use the invariant density by Parry [5], which represents the trajectory of $C_{\beta, \nu}(x)$ for $i = 0, 1, \ldots$ with $x \in [\nu - 1, \nu]$. However, there are two difficulties in using

\footnotetext{1}{In the case of a scale-adjusted $\beta$-map with a full-scale $s$, MSE is multiplied by $s^2(\beta - 1)^2$.}
invariant density for evaluating MSE. One is that the expected applications have the target of L from 12 to 16; these L values are too small to expect for the density function to converge to the invariant measure. The other is that it is also complicated to evaluate the integral on the invariant density by Parry, which is expressed as an infinite sum of $\pm \beta^{-n}s$. This situation motivated us to develop a new MSE analysis method, which is described in the next section.

3. Analysis of segments

In this section, we introduce a new method of analysing $C_{\beta,\nu}^L$ based on segments (see Fig. 3), see also [13]. The graph of $C_{\beta,\nu}^L$ consists of, at most, $2^L$ linear segments, indexed by a bit sequence $b = b_1 \cdots b_L$. We denote the location of each segment by $J_b = [l_b, u_b]$, where $l_b$ and $u_b$ are the lower and upper ends of the projection of the corresponding segment on the y-axis. Segments can be used to calculate the MSE by changing the variable in Eq. (5) as follows:

$$\text{MSE}(C_{\beta,\nu}^L, \theta) = \sum_{b \in \{0, 1\}^L} \int_{l_b}^{u_b} (y - \theta)^2 \frac{dy}{\beta^L}$$

where $J_b = [l_b, u_b]$ and $f_\theta(J_b) = \frac{1}{3}(u_b - \theta)^3 - \frac{1}{3}(l_b - \theta)^3$.

Note that $\sum_{b \in \{0, 1\}^L} \beta^{-L} \mathbb{1}_{[l_b, u_b]}$ can be regarded as $P_{C_{\beta,\nu}^L}(\beta-1) \mathbb{1}_{[0, \frac{1}{\beta-1}]}$, where $P_{C_{\beta,\nu}^L}$ is the Perron-Frobenius operator corresponding to $C_{\beta,\nu}^L$, and $\mathbb{1}_{[0, \frac{1}{\beta-1}]}$ is the indicator function of $[0, \frac{1}{\beta-1}]$. Such a distribution function weakly converges to the invariant density in $L^1([0,1])$ if $\beta > \sqrt{2}$ (see [5]).

Figure 3 shows an example of a fourth-iterated $\beta$-map with $\nu = \frac{\beta-1}{\beta}$. It should be noted that the two segments indexed by 011 and 100 do not appear. This phenomenon emerges if the segments labelled by 011 and 100 in $L = 3$ are included in $[0, \frac{1}{\beta-1}]$ and $[\frac{1}{\beta}, 1]$, respectively. The existence of segments is sensitive to the variation of $\beta$ due to the expanding nature of the $\beta$-map. It makes the analysis of the MSE very difficult. As a solution, we have developed a method to rigorously analyse the $l_b$ and $u_b$ patterns, as shown in the next section. Our analysis shows that the summation in Eq. (6) with at most $2^L$ terms can be reduced to a summation with at most $2L$ terms.

3.1 Segments

We denote the set of segments of $C_{\beta,\nu}^L$ (of positive length) as follows:

Fig. 3. Segments in the map $C_{\beta,\nu}^L$. With $L = 4$, there are 14 segments indexed by $b \in \{0000, \ldots, 1111\}$. Note that segments $J_{0111}$ and $J_{1110}$ have zero length and are not represented in this figure. The lower and upper ends of each segment are given as $l_b$ and $u_b$, respectively. © 2013 IEEE. Reprinted, with permission, from [13, Fig. 3].
Proof. First, remember that the elements of segment $J_b$, that is, $|J_b| = u_b - l_b$. By definition, SEG($i$) has at most $2^i$ elements.

Let the initial segment ($i = 0$) be the input distribution:

$$\text{SEG}(0) = \{J_b\} = \{[0, \frac{1}{2^i}]\}.$$ \hfill (8)

For $i = 1$, we can derive the following:

$$\text{SEG}(1) = \{J_b, J_1\} = \{[0, \nu), [\nu - 1, \frac{1}{2^{i+1}}]\}.$$ \hfill (9)

For a more general $i \geq 2$, the exact set of segments is calculated iteratively.

**Proposition 3.1.** For any bit sequence $b$, the following holds:

$$l_{b0} = \min(\nu, \beta l_b), \quad l_{b1} = \max(\nu, \beta l_b) - 1,$$ \hfill (10)

$$u_{b0} = \min(\nu, \beta u_b), \quad u_{b1} = \max(\nu, \beta u_b) - 1.$$ \hfill (11)

where $b0$ (resp. $b1$) is the concatenation of $b$ and a single bit 0 (resp. 1).

**Proof.** First, remember that $J_{b0} = [l_{b0}, u_{b0})$ and $J_{b1} = [l_{b1}, u_{b1})$ are given as the parts of the images of $J_b$ under $C_{\beta,\nu}$. Using the definition of $C_{\beta,\nu}$, we have the following.

1. When $\beta u_b < \nu$,

$$l_{b0} = C_{\beta,\nu}(l_b) = \beta l_b, \quad u_{b0} = C_{\beta,\nu}(u_b) = \beta u_b.$$ \hfill (12)

Since $J_{b1} = \emptyset$, $l_{b1}$ and $u_{b1}$ can take arbitrary values as long as $l_{b1} = u_{b1}$. Here we set $l_{b1} = u_{b1} = \nu - 1$ for convenience.

2. When $\beta l_b < \nu \leq \beta u_b$,

$$l_{b0} = C_{\beta,\nu}(l_b) = \beta l_b, \quad l_{b1} = \nu - 1,$$ \hfill (13)

$$u_{b0} = \nu, \quad u_{b1} = C_{\beta,\nu}(u_b) = \beta u_b - 1.$$ \hfill (14)

3. When $\beta l_b \geq \nu$,

$$l_{b1} = C_{\beta,\nu}(l_b) = \beta l_b - 1, \quad u_{b1} = C_{\beta,\nu}(u_b) = \beta u_b - 1.$$ \hfill (15)

Since $J_{b0} = \emptyset$, as in the first case, we set $l_{b0} = u_{b0} = \nu$ for convenience.

In each case, we have checked Eqs. (10) and (11), which completes the proof. \hfill \Box

Note that the length of a segment $|J_b|$ may become zero and disappear from the graph of $C_{\beta,\nu}^L(x)$. Such a segment has no effect on the evaluation of the MSE because if $|J_b| = 0$, then $f(J_b) = 0$ and $|J_b| = |J_{b1}| = 0$.

We define this iterative step as the operation $\mathcal{C}$ as follows:

**Definition 3.2.** For each $i \geq 1$, we define an operation $\mathcal{C}$ by

$$\mathcal{C}(J_b) = \{J_{b0}, J_{b1}\},$$ \hfill (16)

$$\mathcal{C}(\text{SEG}(i)) = \bigcup_{J_b \in \text{SEG}(i)} \mathcal{C}(J_b).$$ \hfill (17)

Using this definition, we have

$$\text{SEG}(L + 1) = \mathcal{C}(\text{SEG}(L)).$$ \hfill (18)
3.2 Number of distinct segment locations

Let us investigate how \( \text{SEG}(i) \) varies. In this section we prove the following lemma:

**Lemma 3.3.** For \( L \geq 1 \),

\[
\text{SEG}(L + 1) \supset \text{SEG}(L),
\]
\[
\#\text{SEG}(L + 1) \leq \#\text{SEG}(L) + 2.
\]

where \( \#\text{SEG}(i) \) denotes the number of distinct segment locations in \( \text{SEG}(i) \). Consequently, together with the fact \( \#\text{SEG}(1) = 2 \), we have \( \#\text{SEG}(L) \leq 2L \).

By Lemma 3.3, if we compare \( \text{SEG}(L + 1) \) and \( \text{SEG}(L) \), there exist at most two newly appeared segments. We denote them by \( \mathcal{J}(L + 1), \mathcal{J}(L + 1) \in \text{SEG}(L + 1) \). Their precise definition will be given at the end of this section.

To prove the lemma, we prepare the following sub-lemma, which claims some segments indexed by \( L + 2 \)-bit sequence are the same as other segments indexed by \( L + 1 \)-bit sequence:

**Lemma 3.4.** Let \( L \) be the number of bits, \( b = b_1b_2 \cdots b_L \). Suppose that both \( J_0b_0 \) and \( J_0b_1 \) have positive length. Then,

\[
J_0b_0 \in \text{SEG}(L + 1).
\]

Similarly, suppose that both \( J_1b_0 \) and \( J_1b_1 \) have positive length. Then,

\[
J_1b_1 \in \text{SEG}(L + 1).
\]

**Proof.** First, we prove the following by the induction over the bit sequence \( b \):

\[
l_{0b} = l_b, \quad u_{0b} \leq u_b.
\]

From Eqs. (8)–(9), \( J_0 = [0, \frac{1}{\beta-1}] \) and \( J_0 = [0, \nu] \). Thus for \( b = \emptyset \), Eq. (23) is satisfied.

Suppose that Eq. (23) is satisfied for a given \( b \). Then, by using Proposition 3.1 we have

\[
l_{0b_0} = \min \{ \nu, \beta l_{0b} \} = \min \{ \nu, \beta l_b \} = l_{b_0},
\]
\[
l_{0b_1} = \max \{ \nu, \beta l_{0b} \} - 1 = \max \{ \nu, \beta l_b \} - 1 = l_{b_1},
\]
\[
u_{0b_0} = \min \{ \nu, \beta u_{0b} \} \leq \min \{ \nu, \beta u_b \} = u_{b_0},
\]
\[
u_{0b_1} = \max \{ \nu, \beta u_{0b} \} - 1 \leq \max \{ \nu, \beta u_b \} - 1 = u_{b_1}.
\]

Therefore Eq. (23) is satisfied for \( b' = b_0 \) and \( b_1 \). Thus, by induction of the length of the bit sequence \( b \), Eq. (23) is proved for all \( b \).

Now by using Eq. (23), let us prove the Lemma 3.4. Since both \( J_1b_0 \) and \( J_1b_1 \) have positive length, we have \( \beta l_{0b} < \nu < \beta u_{0b} \). Thus we obtain

\[
J_0b_0 = [\beta l_{0b}, \nu) = [\beta l_b, \nu) = J_{b_0} \in \text{SEG}(L).
\]

Using a similar argument, we have \( J_1b_1 \in \text{SEG}(L) \), which completes the proof. \( \square \)

We shall prove Lemma 3.3 by mathematical induction of \( L \) using Lemma 3.4.

**Proof of Lemma 3.3.** From Eq. (9), it is easy to see that \( \#\text{SEG}(1) = 2 \). We define \( J(1) = J_0 \) and \( \mathcal{J}(1) = J_1 \).

Now, assume that the following is satisfied:

\[
\text{SEG}(L) \subset \text{SEG}(L - 1) \cup \{ \mathcal{J}(L), \mathcal{J}(L) \}.
\]

Let us consider \( \mathcal{C}(\text{SEG}(L)) \). Using the definition,

\[
\mathcal{C}(\text{SEG}(L)) \subset \mathcal{C}(\text{SEG}(L - 1)) \cup \mathcal{C}(\mathcal{J}(L)) \cup \mathcal{C}(\mathcal{J}(L)).
\]
By Eq. (18),

\[ C \left( \text{SEG}(L) \right) = \text{SEG}(L + 1), \quad C \left( \text{SEG}(L - 1) \right) = \text{SEG}(L). \]  

(28)

Using Lemma 3.4 and Proposition 3.1, we see that the set \( C (\mathcal{J}(L)) \setminus \text{SEG}(L) \) contains at most one element of positive length (we represent this element by \( \mathcal{J}(L + 1) \) as this element), and so does \( C (\mathcal{J}(L)) \setminus \text{SEG}(L) \) (we represent this element by \( \mathcal{J}(L + 1) \)). Thus we obtain \( \text{SEG}(L + 1) \subset C (\text{SEG}(L, J(L + 1)). \) \]

\[ \square \]

Note that, in the inductive step from \( \text{SEG}(L) \) to \( \text{SEG}(L + 1) \), the possible value of \( \#(\text{SEG}(L + 1)) - \#(\text{SEG}(L)) \) is 0, 1 or 2: Depending on the value of \( \beta \), \( L \) and \( \nu \), one of the three cases happens when we use Proposition 3.1. The result of the application of this proposition determines the above value.

4. Analysis of the MSE

4.1 Exact MSE for the uniform distribution

Lemma 3.3 shows that the set \( \text{SEG}(i + 1) \) contains at most two new segments of positive length, which do not appear in \( \text{SEG}(i) \). This fact allows us to assign a numerical index \( k \), \( 0 \leq k \leq 2i - 1 \), to the locations of segments. For each \( i \), we define \( J_{2i-2} \) and \( J_{2i-1} \) so that \( J_{2i-2} = J(i) \) and \( J_{2i-1} = J(i) \) hold. Then, for each \( b \in \{0, 1\} \), we define \( k_b \in \{0, \ldots, 2i - 1\} \) such that \( J_b = J_{k_b} \) holds.

This means that, to calculate the MSE for the \( L \)th iteration, we do not have to compute the locations of the \( 2^L \) segments. We only need to count the number, \( n_k^{(L)} \), of segments on the \( k \)th location \( J_k \), for \( k = 0, \ldots, 2L - 1 \). As we will see later, a recurrence formula can be used to calculate \( n_k^{(L)} \). Using the numbers \( n_k^{(L)} \), we can derive the MSE from Eq. (6) as follows:

\[ \text{MSE}(C_{\beta, \nu}, L, \theta) = (\beta - 1)\beta^{-3L} \sum_{k=0}^{2L-1} n_k^{(L)} f_\theta(J_k). \]  

(29)

4.2 Truncating the enumeration of distinct segments

Equation (29) gives the exact MSE, but it is not yet suitable for analysis. For \( L \) which lies in the region of our practical interest, the number of segments \( n_k^{(L)} \) is very sensitive to the values of \( \beta \) and \( \nu \); indeed, the sensitivity grows exponentially with respect to \( L \). This complicates MSE analysis because \( \beta \) is allowed to fluctuate in \( \beta \)-encoders. To solve this problem, we introduce a new way to approximate the MSE based on Lemma 3.3: we call it the \( \beta \)-th-order approximation, which considers only the first 2\( j \) numbers \( n_k^{(i)} \) \( (k = 0, \ldots, 2j - 1) \) of segments for \( i = 1, \ldots, L \).

First, for \( b = b_1 \cdots b_i \in \{0, 1\}^j \), we define

\[ \text{TRUNC}(i, j) = \{ b \in \{0, 1\}^j : \forall i' \leq i \text{ s.t. } k_{b_1 \cdots b_{i'}} < 2j \}, \]

\[ \text{LOST}(i, j) = \{0, 1\}^j \setminus \text{TRUNC}(i, j). \]  

(30)

The set \( \text{TRUNC}(i, j) \) represents the set of segments that is considered under level-\( j \) truncation, that is, the segments whose locations are given by the recursive application of Proposition 3.1 for the first 2\( j \) segments. The set \( \text{LOST}(i, j) \) represents the set of segments that is no longer considered under level-\( j \) truncation. For these lost segments, instead of measuring the length of individual segments precisely, we only calculate their worst contribution based on their average length.

This truncation method has two distinct advantages:

- Given \( \beta \), \( \nu \), and \( j \), we can derive recurrence formulas for \( n_k^{(i,j)} \) and \( n_{\text{lost}}^{(i,j)} \) with respect to \( i \) that considers only 2\( j + 1 \) variables.

- The approximated \( n_k^{(i,j)} \) values are constant for a wider range of \( \beta \) and \( \nu \).

We define the truncated numbers of segments and lost segments as follows:
\begin{align}
n^{(i,j)}_k &= \# \{ b \mid b \in \text{TRUNC}(i,j), J_b = J_k \}, \\
n^{(i,j)}_{\text{lost}} &= \# \text{LOST}(i,j).
\end{align}

This can be calculated by the following recurrence formulas (note that \( \mathbb{I}[\cdot] \) is 1 when the condition is true and 0 otherwise):

\begin{align}
n^{(i,j)}_0 &= n^{(1,j)}_1 = 1, \quad n^{(1,j)}_k = 0 \ (k = 2, \ldots, 2j - 1, \text{lost}), \\
n^{(i,j)}_k &= \sum_{m=0}^{2j-1} n^{(i-1,j)}_m \mathbb{I}[J_k \in \mathcal{C}(J_m)] \ (k = 0, \ldots, 2j - 1), \\
n^{(i,j)}_{\text{lost}} &= 2n^{(i-1,j)}_{\text{lost}} + n^{(i-1,j)}_{2j-2} + n^{(i-1,j)}_{2j-1}.
\end{align}

Let \( h_{\text{lost}} \) be the average length of the segments that belong to the set of lost segments, \( \text{LOST}(L,j) \), defined by

\[ h_{\text{lost}} = \frac{\sum_{b \in \text{LOST}(L,j)} |J_b|}{n^{(L,j)}_{\text{lost}}} = \frac{\beta^L}{\beta - 1} - \frac{\sum_{k=0}^{2j-1} n^{(L,j)}_k |J_k|}{n^{(L,j)}_{\text{lost}}}, \tag{36} \]

for \( n^{(L,j)}_{\text{lost}} \neq 0 \). Jensen’s inequality gives

\[ \sum_{b \in \text{LOST}(L,j)} f_{\theta}(J_b) \leq n^{(L,j)}_{\text{lost}} g(h_{\text{lost}}), \tag{37} \]

for any convex function \( g(x) \) satisfying \( f_{\theta}(J_b) \leq g(|J_b|) \) for all \( l_b \geq \nu - 1 \) and \( u_b \leq \nu \). For the case \( \theta = \theta_0 \), we can use the following function:

\[ g_0(x) = \begin{cases} 
\frac{1}{16}x + \frac{1}{16} & (x \geq \frac{1}{4}), \\
\frac{1}{4}(x - \frac{1}{2})^3 + \frac{1}{24} & (x < \frac{1}{4}).
\end{cases} \tag{38} \]

Then, we have the following upper bounds using level-j truncation:

\[ \text{MSE}(C_{\beta, \nu}, L, \theta_0) \leq \text{MSE}(C_{\beta, \nu}, L, \theta_0, j) \]

\[ \overset{\text{def}}{=} (\beta - 1)\beta^{-3L} \sum_{k=0}^{2j-1} n^{(L,j)}_k f_{\theta_0}(J_k) + n^{(L,j)}_{\text{lost}} g_0(h_{\text{lost}}), \tag{39} \]

These bounds become tighter when \( j \) gets closer to \( L \). When \( j \geq L \), we obtain the non-truncated numbers of segments \( n^{(L,j)}_k = n^{(L,j)}_{\text{lost}} \), and \( n^{(L,j)}_{\text{lost}} \) becomes 0. Hence, when \( j \geq L \), the bound is equal to the exact MSE in Eq. (29).

### 4.3 A simple upper bound for the MSE

Using this truncation, we can give a rigorous proof for the simpler upper bound suggested in our previous work [13] for \( \beta \), \( L \) and \( \nu \) in a practical range for applications.

**Theorem 4.1.** For \( \frac{1 + \sqrt{3}}{2} \leq \beta \leq 2 \), \( 5 \leq L \leq 18 \), and \( \nu = \frac{\beta}{2(\beta - 1)} \),

\[ \text{MSE}(C_{\beta, \frac{\beta}{2(\beta - 1)}}, L, \theta_0) \leq \frac{1}{12} \beta^{-2L}. \tag{40} \]

**Proof.** (Sketch) We divide the range of \( \beta \) into 26 different blocks so that the values of \( n^{(i,8)}_k \) are constant in each block. For each block, we calculate the upper bound of Eq. (39) for \( i = 5, \ldots, 18 \). The complete proof [14] consists of repetition of the arguments for each of \( 26 \times 14 \) cases; an example of the proof for one specific case is presented in Appendix. \( \square \)
5. SQNR of $\beta$-encoders

We have analyzed the quantization error of $\beta$-encoders. However, what we want to see is the impact of the quantization error to the signal-to-noise ratio (SNR), a common criterion for the quality of ADCs. In this section, we give an evaluation of the signal-to-quantization-noise ratio (SQNR) based on our analysis of the MSE.

Suppose that a sinusoidal waveform of frequency $f_{in}$ is sampled and quantized by an ADC. The SQNR is defined, in the frequency domain, as the ratio of the energy of the signal component at frequency $f_{in}$ to the total energy of quantization noise component. Since the range of the $\beta$-encoder is $[0, \frac{1}{\beta - 1}]$, the energy of the input signal for one period (which we denote by $S$) is equal to $(\beta - 1)^{-2}/8$.

Here we approximate the noise component $N$ by $\text{MSE}(C_{\beta,\nu}, L, \theta)$, that is, the quantization MSE. This means that we approximate the distribution of the sampled sinusoidal waves with the uniform one over $[0, \frac{1}{\beta - 1}]$. This results in the following formula:

$$\text{SQNR}(\theta) = 10 \log_{10} \frac{S}{N} \approx 10 \log_{10} \frac{(\beta - 1)^{-2}/8}{\text{MSE}(C_{\beta,\nu}, L, \theta)} \text{ (dB)}. \quad (41)$$

Using Eq. (40), we get the approximated lower bounds of the SQNR as follows:

$$\text{SQNR}(\theta_0) \geq 20L \log_{10} \beta - 20 \log_{10}(\beta - 1) + 1.76 \text{ (dB)}. \quad (42)$$

For a scale-adjusted $\beta$ map with scale $s$, both the signal and noise components are multiplied by $s^2(\beta - 1)^2$. Thus, Eq. (41)–(42) hold for any $s$.

6. Simulation results

Figures 4 and 6 show a comparison of the derived MSE values and the results of numerical simulations for $L = 18$ and $\nu = \frac{\beta}{\sqrt{\beta - 1}}$. The curve labeled as “Exact” shows the value of Eq. (29) calculated numerically. The curve labeled as “Simulation” is the result of the calculation of Eq. (5) for each $\beta$ value using MATLAB, evaluating the integral using $2^{18}$ uniformly distributed points. Since the simulation results perfectly match the exact MSE value, they are not distinguishable in the plot. The curve labeled “level-8 trunc.” shows the truncated MSE value $\hat{\text{MSE}}(C_{\beta,\nu}, L, \theta, j)$ (see Eq. (39)) letting $L = 18$ and $j = 8$. For the range $1 + \sqrt{5}/2 \leq \beta < 2$, the MSE based on level-8 truncation is better than the simple bounds in Eqs. (40) though it is not the case for some of the other parameters.

Figures 5 and 7 show a comparison of the SQNR values with a MATLAB-based numerical simulation of a sinusoidal wave input consisting of $2^{18}$ samples. The SNR calculated from the simulation is slightly worse than the SQNR calculated from the exact MSE and Eq. (41), but it is better than the simple bound in Eq. (42) for $1 + \sqrt{5}/2 < \beta < 2$. Moreover, the difference between the simulation and the simple bound is less than 0.8 dB. This means that Eq. (42) can be used as a good approximation of the ideal SQNR value of the $\beta$-encoder.

![Fig. 4. Comparison of MSES ($L = 18$). For the explanation of each curve, see Section 6.](image)

![Fig. 5. Comparison of SQNRs (dB), $L = 18$.](image)
Fig. 6. Comparison of the MSE ratios (simple bound = 1, $L = 18$).

Fig. 7. Comparison of the SQNR differences (simple bound = 0, $L = 18$).

Fig. 8. Ratio of MSEs calculated by level-8 truncation and the simple bound for a wider range of $\beta$ and $L$. Values are encoded with $\nu = \frac{\beta}{2(\beta - 1)}$. The red rectangular region denotes the condition in Theorem 4.1.

Fig. 9. Ratio of MSEs calculated by numerical simulation to the simple bound for a wider range of $\beta$ and $L$. Values are encoded with $\nu = \frac{\beta}{2(\beta - 1)}$. The red rectangular region denotes the condition in Theorem 4.1.

Figures 8 and 9 show the effectiveness of the simple bound for a wider range of the parameters. We can see in Fig. 8 that the red region (the parameters satisfying the preconditions in Theorem 4.1) touches the boundary of the blue region, where $\text{MSE}(C_{\beta, \nu}, L, \theta_0, 8)$ (the level-8 truncated MSE, see Eq. (39)) is smaller than the simple bound $\beta^{-2L}/12$. However, the simulation result in Fig. 9 suggests that Eq. (40) actually holds for a much wider region such as $L > 18$. We may be able to prove it by increasing the level of the truncation, but it comes at a cost of exponential increase of the number of cases in the proof.

Figure 10 shows the simulation result with varying $\nu$ in the region $1 \leq \nu \leq \frac{1}{\sqrt{2} - 1}$, in which the $\beta$-map is defined. Equation (40) is true if we set $\nu$ be close to $\nu^{*} = \frac{\beta}{2(\beta - 1)}$ (the midpoint of the region), while as $\nu$ gets close to the boundary of the region (the so-called “greedy” or “lazy” $\beta$-map), the MSE increases and violates Eq. (40) for some $\beta$. The red region in the figure suggests that Eq. (40) is true even if $\nu$ has some fluctuations.

7. Conclusion
We have proposed a method for providing a rigorous analysis of the MSE of quantization error in ADCs based on $\beta$-map. Using an analysis based on segments of nested $\beta$-maps, we have given a rigorous proof that the MSE is bounded from above by $\frac{1}{12}\beta^{-2L}$ in a practical range of $L$ and $\beta$. Such an MSE analysis guarantees the quality of $\beta$-encoders and leads to a useful SQNR evaluation, which
Fig. 10. Comparison of the simulated MSE ratios (simple bound = 1) for a wider range of \( \beta \) and \( \nu \), where \( \nu^* = \frac{\beta}{2(\beta-1)} \), \( L = 18 \). The red region denotes \( \nu = \nu^* \times (1 \pm 0.05) \) and \( \frac{1+\sqrt{5}}{2} \leq \beta \leq 2 \).

is an indispensable step for determining \( L \) and \( \beta \) in the design of \( \beta \)-encoder circuits to meet demands of practical applications.

**Acknowledgments**

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

**Appendix**

Here we give the proof of Theorem 4.1 for one specific case (1.6181 < \( \beta \) < 1.6483 and \( L = 18 \)). The whole proof consists 26 \( \times \) 14 cases, but the other cases are proved in the same manner (see also Fig. A-1) We also have the whole proof [14], but it is generated by a program and too tedious to be put here.

We calculate the ratio of the truncated MSE and the simple upper bound,

\[
R_{L,\theta_0}(\beta) = \frac{\text{MSE}(C_{\beta,\nu}, L, \theta_0, j)}{\frac{1}{2} \beta - 2L},
\]

and show that \( R_{L,\theta}(\beta) < 1 \) for each case. Since the exact MSE is always bounded by the truncated MSE (Eq. (39)), this completes the proof.

The boundaries of each block of \( \beta \) are actually the roots of two algebraic equations. The lower bound 1.6181 is the larger root of \( \beta^2 - \beta - 1 = 0 \), which is \( \frac{\sqrt{5}}{2} + \frac{1}{2} \). The upper bound 1.6483 is the

**Fig. A-1.** Plot of \( R_{L,\theta_0}(\beta) \) for \( \frac{1+\sqrt{5}}{2} \leq \beta \leq 2 \) and \( 5 \leq L \leq 18 \).
real root of $\beta^5 - \beta^4 - 2\beta^2 + \beta - 1 = 0$. For this range of $\beta$, we can calculate the formulas of segment locations $(l_k), (u_k)$ and the number of populations $n^{(18,8)}_k$ as follows:

\[
\begin{pmatrix}
  l_0 \\ l_1 \\ \vdots \\ l_{15}
\end{pmatrix} = \frac{1}{2(\beta - 1)} \begin{pmatrix}
  0 \\ 2 - \beta \\ 2 - \beta \\ 2\beta - \beta^2 \\ 2 - \beta \\ 2\beta - \beta^2 \\ 2\beta - \beta^2 \\ -\beta^5 + 2\beta^4 - 2\beta + 2 \\ 2\beta^2 - \beta^3 \\ -\beta^6 + 2\beta^5 - 2\beta^2 + 2 \\ 2 - \beta \\ -\beta^7 + 2\beta^6 - 2\beta^3 + 2\beta \\ 2\beta - \beta^2 \\ -\beta^8 + 2\beta^7 - 2\beta^4 + 2\beta^2 - 2\beta + 2
\end{pmatrix}
\]

\[
\begin{pmatrix}
  u_0 \\ u_1 \\ \vdots \\ u_{15}
\end{pmatrix} = \frac{1}{2(\beta - 1)} \begin{pmatrix}
  \beta \\ 2 - \beta + 2 \\ \beta \\ 2 - 2\beta + 2 \\ \beta^3 - 2\beta^2 + 2 \\ \beta \\ 2 - \beta + 2 \\ \beta^5 - 2\beta^4 + 2\beta \\ \beta^2 - \beta^2 + 2 \\ \beta^6 - 2\beta^5 + 2\beta^2 \\ \beta^3 - 2\beta^2 + 2 \\ \beta^7 - 2\beta^6 + 2\beta^3 - 2\beta + 2 \\ \beta^8 - 2\beta^7 + 2\beta^4 - 2\beta^2 + 2\beta \\ \beta^2 - 2\beta + 2
\end{pmatrix}
\]

\[
\begin{pmatrix}
  n^{(18,8)}_0 \\ n^{(18,8)}_1 \\ \vdots \\ n^{(18,8)}_{15} \\
  n^{(18,8)}_{\text{lost}}
\end{pmatrix} = \begin{pmatrix}
  1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1
\end{pmatrix}
\]

In this range, $n_{\text{lost}} < 1/4$. Hence, combining Eqs. (A-1)–(A-2) and Eq. (39) with Eq. (7) yields the following:

\[
R_{18,\theta_0}(\beta) = \frac{1}{3888 + \beta^{12}(\beta - 1)^2} (\beta^{54} - 900\beta^{44} + 336\beta^{43} + 549\beta^{42} + 900\beta^{41} - 336\beta^{40} - 552\beta^{39} - 11664\beta^{36} + 11664\beta^{36} + 270000\beta^{34} - 201600\beta^{33} - 291768\beta^{32} - 417024\beta^{31} + 900\beta^{41} - 336\beta^{40} - 552\beta^{39} - 11664\beta^{36} + 11664\beta^{36} + 270000\beta^{34} - 201600\beta^{33} - 291768\beta^{32} - 417024\beta^{31} + 503667\beta^{30} + 585336\beta^{29} + 23376\beta^{28} + 6594768\beta^{27} - 9904704\beta^{26} - 1532640\beta^{25} + 4505335392\beta^{24} - 271669928064\beta^{23} + 54459358416\beta^{22} - 28861612352\beta^{21} - 17120085444\beta^{20} - 20510713296\beta^{19} + 34678912551\beta^{18} + 107496632580\beta^{17} - 208144893312\beta^{16} + 110996021784\beta^{15} - 8411418048\beta^{14} + 91468794864\beta^{13} - 142465482064\beta^{12} - 212204613504\beta^{11} + 501492475392\beta^{10} + 50236497152\beta^{9} - 694609862400\beta^{8} - 534758083820\beta^{7} - 1558799335488\beta^{6} - 48995742128\beta^{5} + 1452367314432\beta^{4} - 1360488960000\beta^{3} + 423384164352\beta^{2})
\]

Let us investigate the behavior of the function $R_{18,\theta_0}(\beta)$ for $1.6181 < \beta < 1.6483$. The boundary values of $\beta$ are $R_{18,\theta_0}(1.6181) = 0.8295$ and $R_{18,\theta_0}(1.6483) = 0.99764$. In this range the formula
\[ \frac{dR_{18, \theta_0}}{d\beta} = 0 \] has a unique root at \( \beta = 1.6457 \), where the function \( R_{18, \theta_0}(\beta) \) has a local maximum with the value \( R_{18, \theta_0}(1.6457) = 0.99887 \). Therefore \( R_{18, \theta_0}(\beta) \leq 0.99887 \) is proved in this case.

References

[1] I. Daubechies, R.A. DeVore, C.S. Güntürk, and V.A. Vaishampayan, “Beta expansions: A new approach to digitally corrected A/D conversions,” In Proc. IEEE Int. Symp. Circuits Syst., vol. 2, pp. 784–787, 2002.

[2] I. Daubechies, R.A. DeVore, C.S. Güntürk, and V.A. Vaishampayan, “A/D conversion with imperfect quantizers,” IEEE Trans. Inform. Theory, vol. 52, pp. 874–885, 2006.

[3] A. Rényi, “Representations for real numbers and their ergodic properties,” Acta Math. Hungar., vol. 8, pp. 477–493, 1957.

[4] W. Parry, “On the \( \beta \)-expansions of real numbers,” Acta Math. Acad. Sci. Hungar., vol. 11, pp. 401–416, 1960.

[5] W. Parry, “Representations for real numbers,” Acta Math. Acad. Sci. Hungar., vol. 15, pp. 95–105, 1964.

[6] T. Kohda, Y. Horio, and K. Aihara, “\( \beta \)-expansion attractors observed in A/D converters,” AIP Chaos, An Interdisciplinary Journal of Nonlinear Science, vol. 22, no. 4, 047512-1-16, 2012.

[7] W. Kester, ed., The Data Conversion Handbook, Elsevier: Newnes, 2005.

[8] H. San, T. Kato, T. Maruyama, K. Aihara, and M. Hotta, “Non-binary pipeline analog-to-digital converter based on \( \beta \)-expansion,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, vol. E96-A, no. 2, pp. 415–421, 2013.

[9] R. Suzuki, T. Maruyama, H. San, K. Aihara, and M. Hotta, “Robust cyclic-ADC architecture based on \( \beta \)-expansion,” IEICE Trans. on Electronics, vol. E96-C, no. 4, pp. 553–559, 2013.

[10] I. Daubechies and Ö. Yilmaz, “Robust and practical analog-to-digital conversion with exponential precision,” IEEE Trans. Inform. Theory, vol. 52, pp. 3533–3545, 2006.

[11] R. Ward, “On robustness properties of beta encoders and golden ratio encoders,” IEEE Trans. Inform. Theory, vol. 54, no. 9, pp. 4324–4334, 2006.

[12] T. Kohda, Y. Horio, Y. Takahashi, and K. Aihara, “Beta encoders: Symbolic dynamics and electronic implementation,” International Journal of Bifurcation and Chaos, vol. 22, no. 9, 1230031, 2012.

[13] T. Makino, Y. Iwata, Y. Jitsumatsu, M. Hotta, H. San, and K. Aihara, “Rigorous analysis of quantization error of an A/D converter based on \( \beta \)-map,” In Proc. of 2013 IEEE Int. Symp. on Circuits and Systems (ISCAS2013), pp. 369–372, 2013.

[14] T. Makino, “Proving the upper bound of quantization MSE of the \( \beta \)-encoder,” available from http://www.sat.t.u-tokyo.ac.jp/~mak/betaproof.pdf, 2014.

[15] K. Dajani and C. Kraaikamp, “From greedy to lazy expansions and their driving dynamics,” Exp. Math., vol. 20, pp. 315–327, 2002.