論文 / 著書情報
Article / Book Information

Title	Stable Single-Bit Noise-Shaping Quantizer Based on Sigma-Delta Modulation and Recursive Data Coding into Pre-Optimized Binary Vectors
Authors	Mitsuhiko Yagyu, Akinori Nishihara
Citation	IEICE Trans. Fundamentals., Vol. E85-A, No. 8, pp. 1781-1788
Pub. date	2002,
URL	http://search.ieice.org/
Copyright	(c) 2002 Institute of Electronics, Information and Communication Engineers

PAPER *Special Section on Digital Signal Processing*

Stable Single-Bit Noise-Shaping Quantizer Based on ΣΔ Modulation and Successive Data Coding into Pre-Optimized Binary Vectors

Mitsuhiko YAGYU[†a)] ***and*** **Akinori NISHIHARA**[††], ***Regular Members***

SUMMARY This paper presents data coding techniques for a stable single-bit noise-shaping quantizer, which has a cascade structure of a multi-bit ΣΔ modulator and a binary interpolator. The binary interpolator chooses a pre-optimized binary vector for each input sample and successively generates the chosen binary vector as an output bit stream. The binary vectors can have different lengths. The paper also proposes two methods to evaluate and bound output errors of a binary interpolator. A multi-bit ΣΔ modulator is designed to cause no overload for all possible input signals whose amplitudes are bounded to a specified level, and thus the ΣΔ modulator rigorously guarantees the stability condition. In design examples, we have evaluated Signal-to-Noise and Distortion Ratios (SNDRs) and noise spectra and then confirmed that our stable quantizers can sharply shape output noise spectra.

key words: *stability, single-bit quantizer, noise-shaping, ΣΔ modulator, data coding*

1. Introduction

For oversampling A-D and D-A conversions that utilize noise-shaping techniques, a high-order ΣΔ modulator can be used to shape quantization noise spectra and can significantly reduce the in-band noise power even with a moderate oversampling ratio [1]. Specifically, a D-A converter (DAC) with a single-bit ΣΔ modulator can perform perfect linearity, but high-order single-bit modulators often suffer from the problem of instability. By employing a multi-bit ΣΔ modulator, the problem of instability may be alleviated. In this case, since linearity of a multi-bit DAC is sensitive to analog component mismatches, the mismatch shaping dynamic element matching (DEM) technique (e.g. [2], [3]), which can compensate the effect of the mismatches, is often applied to the multi-bit ΣΔ modulator. The multi-bit ΣΔ modulation together with the mismatch shaping DEM technique would reduce the quantization noise and an error due to the analog component mismatches in a band of interest at a moderate oversampling ratio.

Manuscript received December 10, 2001.
Final manuscript received April 19, 2002.

[†]The author is with the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo, 184-8588 Japan.
[††]The author is with the Center for Research and Development of Educational Technology, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
a) E-mail: myagyu@cc.tuat.ac.jp

Another approach is to use a single-bit DAC with a stabilized single-bit digital quantizer which performs a sharp noise-shaping. The technique to employ a multi-bit ΣΔ modulator with the mismatch shaping DEM would reduce the in-band error due to the mismatches, but anyway it cannot perfectly compensate the error, and the compensation depends on statistics of magnitudes of the mismatches. On the other side, a single-bit DAC with a single-bit digital quantizer still has an advantage that matching the amplitudes of quantization levels is not an issue, and a single-bit DAC can inherently perform perfect linearity. However, the serious problem to stabilize high-order single-bit ΣΔ modulators and bound their output errors has not completely been solved. Although many criteria and concepts of stability conditions have been proposed (e.g. [4], [5]), the problem how to guarantee the stability has still been open from theoretical point of view. In other words, with a stable single-bit quantizer, we could perfectly fix the issue due to the analog component mismatch by employing a single-bit DAC, although single-bit high-order noise-shaping quantizers with rigorous stability have not been proposed.

In this paper, we propose stable single-bit digital quantizers which have a cascade structure of a multi-bit ΣΔ modulator and a binary interpolator as shown in Fig. 1. In order to guarantee rigorous stability of the multi-bit ΣΔ modulator from theoretical point of view, we introduce the no-overload condition [1], [5], in which l_1 norm of coefficients of the noise transfer function $1 - H_{fb}(z)$ is limited to a certain level, and then the coefficients are optimized to minimize the in-band noise. Under the no-overload condition, any overload cannot occur at the roundoff operator so that the roundoff error is strictly bounded for an arbitrary in-

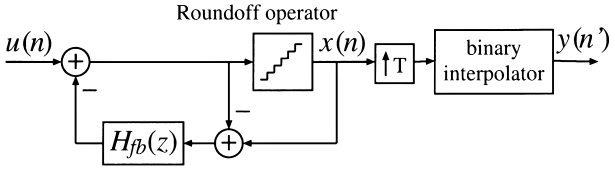


Fig. 1 A block diagram of the proposed digital quantizer. The quantizer has a cascade structure of a multi-bit ΣΔ modulator and binary interpolator.

put signal whose maximum amplitude is bounded to a specified level. This no-overload condition is a sufficient condition for the stability and may be conservative, but we can analytically see that it rigorously guarantees the stability. The stable multi-bit $\Sigma\Delta$ modulator coarsely quantizes its input signal to a train of a sample having several bits. The output of the stable multi-bit $\Sigma\Delta$ modulator is then up-sampled by a factor T, and the coarsely quantized sequence is interpolated by a binary interpolator. The binary interpolator maps each input sample into a corresponding binary vector that is pre-optimized to have minimum in-band error, and then the binary interpolator generates a sequence of binary vectors as a bit stream to be D-A converted by a single-bit DAC. Such a binary interpolator is a nonlinear system, and thus its output error, namely, its quantization noise spectrum cannot explicitly be expressed as an error of a linear time-invariant system. So we propose two methods to analyze its maximum output error spectrum (MOES) and mean squared output error spectrum (MSOES) for all possible input signals. In design examples, we first design a binary interpolator and analyze its output errors by using the above methods. Then it is confirmed that actual output errors can be tightly upper-bounded by using the methods. We design and optimize a stable single-bit digital quantizer and then demonstrate the effectiveness of its performance.

2. Binary Interpolator

2.1 Principle of Interpolation Algorithm

Figure 2 depicts a relation between an input and output signal of a binary interpolator for $T = 6$ as an example. The binary interpolator chooses a proper binary vector corresponding to each input sample $x(n)$ and then generates a sequence of the chosen binary vectors as an output binary bit stream of the binary interpolator. The binary vectors can have different lengths. As shown in the figure, suppose that the two input samples $x(n)$ and $x(n+2)$ have an equal value, but a vector corresponding to $x(n)$ must have a delay of -4, and the delay of a vector chosen for $x(n+2)$ must be -3 in this example. Thus a binary vector needs to be chosen by taking account of not only the value of each input sample but also the delay that is successively calculated for each input sample. Note that the binary interpolator shown in Fig. 2 is assumed to be non-causal for convenience. To implement the binary interpolator, a system delay is introduced so as to cancel the largest negative delay for all possible binary vectors so that the binary interpolator can be causal.

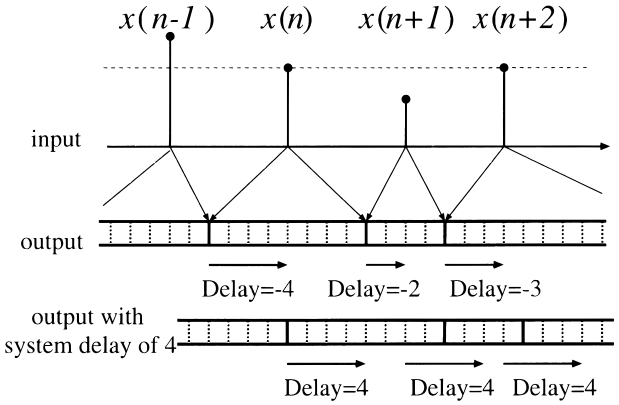


Fig. 2 An example of a relation between input and output signals of a binary interpolator with $T = 6$.

2.2 Output Error of Binary Interpolators

Let a binary vector chosen for $x(n)$ be $v[n, m]$, where m is the time index for signals up-sampled by T and also is a pointer to each element of the vector. Here we consider that the binary interpolator has a system delay of zero for convenience and, namely, it is implemented as a delay-free system. Then assume that the vector $v[n, m]$ having a length of $l_v(n)$ has a negative delay of $-\tau(n)$ as shown in Fig. 2. Then we define $v[n, m]$ as a binary vector;

$$v[n,m]=\begin{cases} \pm 0.5, & -\tau(n) \le m < l_v(n) - \tau(n), \\ 0.0, & \text{otherwise} \end{cases} . \quad (1)$$

An output binary signal $y(n')$ shown in Fig. 1 is a train of binary vectors $v[n, n' - nT]$ for $-\infty < n < \infty$, which is chosen at time nT, and thus written as

$$y(n') = \sum_{n=-\infty}^{\infty} v[n, n' - nT]. \quad (2)$$

Then its Fourier transform is written as

$$Y(e^{j\omega}) = \sum_{n=-\infty}^{\infty} V[n, e^{j\omega}]e^{-jnT\omega}, \quad (3)$$

where $V[n, e^{j\omega}]$ is the Fourier transform of the binary vector chosen at time nT. Also $e^{-jnT\omega}$ in Eq. (3) can be regarded as the time-shift operator in the Fourier transform for the binary vector chosen at time nT. Here let a specified band of interest be $|\omega| \le \omega_D < \pi/T$. We assume that an ideal output signal of the binary interpolator should be identical to the input signal $x(n)$ in the band of interest, because the binary interpolator has been assumed to be a delay-free system. Hence the ideal output signal can be defined as

$$D(\omega) \sum_{n=-\infty}^{\infty} x(n)e^{-jnT\omega}, \quad (4)$$

where

$$D(\omega)=\begin{cases} 1, & |\omega| \le \omega_D, \\ 0, & \omega_D < |\omega| < \pi \end{cases} . \quad (5)$$

In this paper, the output error $R_o(e^{j\omega})$ of the binary interpolator is defined as the difference between the output bit stream and ideal output signal, and by using Eqs. (3) and (4) it can be written as

$$R_o(e^{j\omega}) = \sum_{n=-\infty}^{\infty} \left\{V[n, e^{j\omega}] - x(n)D(\omega)\right\} e^{-jnT\omega} \quad (6)$$

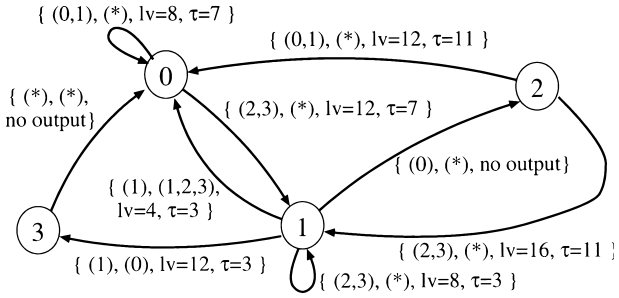
$$= \sum_{n=-\infty}^{\infty} R[n, e^{j\omega}]e^{-jnT\omega}, \quad (7)$$

where $R[n, e^{j\omega}]$ is a function of ω and then can be regarded as the error response of the binary vector chosen at time nT in the frequency domain. From Eq. (7), if ω is fixed at a frequency ω_a, we find that the output error $R_o(e^{j\omega_a})$ at the frequency ω_a can be written as the Fourier transform of a time sequence of complex numbers $R[n, e^{j\omega_a}]$ for $n = \cdots, -1, 0, 1, \cdots$. We refer to this time sequence as Output Error Estimation Sequence (OEES) at the frequency ω_a [6].

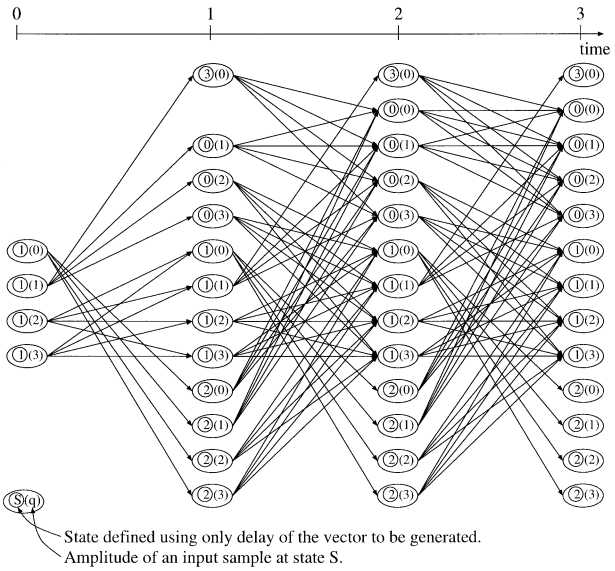
The operation of a binary interpolator can be described by using a state transition diagram. Figure 3 shows a simple example of such state transition. We assume that an input sample to the binary interpolator can take seven kinds of values; 0, ±1, ±2 and ±3 in this example. In Fig. 3(a), each state transition is conditioned by an indication such as $\{(q_1, q_2, q_3), (q_4, q_5), l_v, \tau\}$. Suppose a state transition from state S at time nT to S' at time $(n+1)T$ is conditioned by that indication. The indication means that, if $x(n) = q_1$, q_2 or q_3 and $x(n+1) = q_4$ or q_5, the binary interpolator chooses the binary vector which is pre-optimized for the two amplitudes of $x(n)$ and $x(n+1)$ and has a length of l_v and a negative delay of $-\tau$. Next the binary interpolator generates the chosen vector as an output vector at time nT, and then the state of the binary interpolator changes to S' at time $(n+1)T$.

For a negative value of an input sample $x(n)$, the binary interpolator first chooses the binary vector $v(n, m)$ for $|x(n)|$ and $|x(n+1)|$, which has the proper l_v and $-\tau$, and generates a vector $-v(n, m)$. If "no output" is indicated, the binary interpolator generates no output. The asterisk "$*$" stands for "don't care." We need to pre-determine the length l_v and negative delay $-\tau$ such that the output signal is a sequence of only the two values ±0.5 even with any state transition. Figure 3(b) shows the trellis diagram which is equivalent to the state transition diagram. Successive state transitions caused by any input signals can be interpreted by Fig. 3(b).

Here we present a method to analyze output errors of a binary interpolator. The method is based on [6]. Let the number of samples and quantization levels of $x(n)$ be P and $\pm q_i$, for $i = 1, \cdots, L$, respectively. Here we consider a set of all possible input signals with P and q_i in order to evaluate output errors at an arbitrary



(a) A state transition diagram defined using only delay of each vector to be generated.



(b) The trellis diagram derived from combinations of an input amplitude and the delay of each vector to be generated.

Fig. 3 A simple example of state transition of a binary interpolator with four states and a system delay τ of 3.

frequency ω_a and then define $\Psi(P, L)$ as it. Then with all possible signals of $\Psi(P, L)$, we define the maximum output error at the frequency ω_a as

$$R_{worst}(\omega_a, P) = \frac{1}{P} \max_{\substack{x(n) \\ \in \Psi(P,L)}} \left| \sum_{n=0}^{P-1} R[n, e^{j\omega_a}]e^{-jn\omega_a} \right|. \quad (8)$$

Equation (8) can be upper-bounded as

$$R_{worst}(\omega_a, P) \le \frac{1}{P} \max_{\substack{x(n) \\ \in \Psi(P,L)}} \sum_{n=0}^{P-1} \left|R[n, e^{j\omega_a}]\right|. \quad (9)$$

The upper bound (9) is calculated as the sum of error amplitudes, which are evaluated at the frequency ω_a, of binary vectors chosen for input samples. We now consider an algorithm to efficiently calculate the upper bound (9). An arbitrary input signal of $\Psi(P, L)$ is associated with a unique sequence of states as shown in Fig. 3(b). However generally a sequence of states defined by the trellis diagram can be associated with

several input signals, because a branch of the trellis diagram can indicate a state transition which is caused by several different input values. For example, the state transition from state $(S, q) = (1, 1)$ at time n to state $(0, 2)$ at time $n + 1$ occurs, if the input sample has a value of ± 2 at time $n + 1$. In such a case, errors of the two binary vectors are compared at frequency ω_a, and larger error value is referred to as the cost of the branch. In the same way, we can determine a cost of each branch. Then by using the Viterbi algorithm [7], we can find a path which has maximum sum of costs, and the maximum sum divided by P corresponds to the upper bound. The trellis diagram has a periodic and regular structure. Evaluating only certain short length of the trellis diagram by the Viterbi algorithm, sometimes we can efficiently calculate the upper bound for the case of large P.

Next we assume that input signals are statistically defined to be a strictly white process with a probability density function given as

$$\sum_{i=1}^{L} \{p(q_i)\delta(x - q_i) + p(-q_i)\delta(x + q_i)\}, \tag{10}$$

where $p(q_i)$ is the probability that the input signal $x(n)$ takes the value of q_i. Then $R[n, e^{j\omega_a}]$, where the frequency ω_a is fixed, is a time sequence of complex numbers for $n = 0, \cdots, P-1$ and then becomes a statistical process. We define the mean squared output error at ω_a as

$$R_{mse}(\omega_a, P) = E\left[\frac{1}{P}\left|\sum_{n=0}^{P-1} R[n, e^{j\omega_a}]e^{-jn\omega_a}\right|^2\right]. \tag{11}$$

For example, in Fig. 3(b) note that if the input signal with large P is strictly stationary, the probability that state (S, q) is changed into $(0, 2)$ at time n can be written as

$$\{p(2) + p(-2)\} \Pr\{(S, q) = (0, 0),\ (0, 1),\ (1, 1),\ (2, 0),\ (2, 1),\ \text{or } (3, 0) \text{ at time } n - 1.\}. \tag{12}$$

In the same way, the probability that state is changed into an arbitrary state at time n can be written with the probability density function of the input signal and trellis structure. Also a state transition between two arbitrary states can be described by using a state transition matrix Π [8]. Then the joint probability density of the statistical sequence $R[n, e^{j\omega_a}]$, where n is the time index of this process, can be easily calculated so that the average auto-correlation function of $R[n, e^{j\omega_a}]$ can be obtained. Then its Fourier transform at ω_a gives $R_{mse}(\omega_a)$.

2.3 Minimization of Output Error at Specified Frequency Band

From Eq. (9), we see that the maximum output error at a frequency ω_a for all possible input signals can be upper-bounded by using the error value of the complex function $V[n, e^{j\omega}]$, which is Fourier transform of the chosen binary vector, at the frequency ω_a. Then the error value is defined as the difference between the two complex numbers $V[n, e^{j\omega_a}]$ and $x(n)D(\omega)$ from Eq. (6). This implies that if the Fourier transforms of all possible chosen binary vectors have no error at a specified frequency, the output signal of the binary interpolator also has no error at the frequency for arbitrary input signal. Equation (9) also means that the peak output error at the frequency ω_a in the worst case can be reduced by minimizing the error value of $V[n, e^{j\omega_a}]$. In this paper, we optimize the binary vectors to be chosen for each input sample and then minimize the peak error of those vectors in a band of interest.

2.4 Optimization of Binary Vectors

Remember that a binary vector to be optimized has been assumed to have a length l_v and a negative delay of $-\tau$. We rewrite a binary vector for an input amplitude of q_i as $v_i(m)$ and its Fourier transform as $V_i(e^{j\omega})$, respectively. In this paper, $V_i(e^{j\omega})$ is optimized to have no error at DC and a maximally flat response. $V_i(e^{j\omega})$ is written as

$$V_i(e^{j\omega}) = \sum_{m=-\tau}^{l_v-\tau-1} v_i(m)\cos m\omega - j\sum_{m=-\tau}^{l_v-\tau-1} v_i(m)\sin m\omega. \tag{13}$$

so that we obtain

$$\frac{d^k}{d\omega^k}V_i(e^{j\omega}) = \sum_{m=-\tau}^{l_v-\tau-1} m^k v_i(m)\cos(m\omega + k\pi/2) - j\sum_{m=-\tau}^{l_v-\tau-1} m^k v_i(m)\sin(m\omega + k\pi/2) \tag{14}$$

The Fourier transform of the optimized binary vector needs to approximate q_i, which is the Fourier transform of the input sample in the band of interest. As discussed in Sect. 2.3, if each optimized binary vector has no error at DC, that is, meets

$$\sum_{m=-\tau}^{l_v-\tau-1} v_i(m) = q_i, \tag{15}$$

then any output signals of the binary interpolator do not have any errors at DC. From Eq. (14), the condition that the Fourier transform $V_i(e^{j\omega})$ of the binary vector approximates q_i at DC with $(K + 1)$-th order flatness is written as

$$\sum_{m=-\tau}^{l_v-\tau-1} m^k v_i(m) = 0 \tag{16}$$

for $k = 0, \cdots, K$. Namely, to reduce the output error with $(K+1)$-th order flatness, a problem to optimize $v_i(m)$ for $m = -\tau, \cdots, l_v - \tau - 1$ can be formulated as

$$\begin{aligned} \text{Minimize} \quad & \left| \sum_{m=-\tau}^{l_v-\tau-1} m^K v_i(m) \right| \\ \text{Subject to} \quad & \sum_{m=-\tau}^{l_v-\tau-1} v_i(m) = q_i, \\ & \sum_{m=-\tau}^{l_v-\tau-1} m^k v_i(m) = 0, \\ & \text{and } v_i(m) = \pm 0.5 \\ & \text{for } m = -\tau, \cdots, l_v - \tau - 1 \\ & \text{and } k = 1, \cdots, K-1. \end{aligned} \tag{17}$$

In our design algorithm, if the objective function becomes zero with the optimum solution, i.e., an optimum solution performs the perfect $(K+1)$-th order flatness, then K is incremented, and new optimization problem for the incremented K is solved. By this iterative procedure, we maximize K, namely, improve the flatness. To find possible state transitions, the above optimization problem is solved for $\tau = 0$, 0.5, 1.0, $\cdots$, $l_v - 1$ and a wide range of l_v by exhaustive search. Then we heuristically find several possible combinations of state transitions.

3. Multi-Bit ΣΔ Modulators

In Fig. 1, we assume that the roundoff operator rounds off its input to the nearest integer. Let an impulse responses of the noise transfer function be $h_{NTF}(m)$, $m = 0, \cdots, T_{NTF} - 1$, where $h_{NTF}(0)$ is 1 from Fig. 1. Then the impulse response $h_{fb}(m)$ of $H_{fb}(z)$ is given as $h_{fb}(0) = 0$ and $h_{fb}(m) = -h_{NTF}(m)$ for $m = 1, \cdots, T_{NTF} - 1$. Let a quantization level of the roundoff operator be 0, ± 1, ± 2, $\cdots$, $\pm Q$, where Q is a positive integer, and assume that the roundoff operator carries out the unbiased roundoff. Then if the amplitude of its input is upper-bounded by $Q + 0.5$, which we call the no-overload condition in this paper, the roundoff error is upper-bounded by 0.5. Here let the allowable maximum amplitude of $u(n)$ be u_{max}. Then if $||h_{fb}(m)||_1 \times 0.5 + u_{max}$, which is the maximum of inputs to the roundoff operator, is less than $Q + 0.5$ and all the delay elements of $H_{fb}(z)$ is initialized to be zero when started up, the amplitude of any roundoff errors are strictly upper-bounded by 0.5, and then the multi-bit ΣΔ modulator is stable [1], [5].

Next, we formulate a problem to optimize an impulse response $h_{NTF}(m)$ as

$$\begin{aligned} \text{Minimize} \quad & \left| \sum_{m=0}^{T_{NTF}-1} m^K h_{NTF}(m) \right| \\ \text{Subject to} \quad & \sum_{m=0}^{T_{NTF}-1} m^k h_{NTF}(m) = 0, \\ & h_{NTF}(0) = 1, \\ & \sum_{m=1}^{T_{NTF}-1} |h_{NTF}(m)| \le c, \\ & \text{and } k = 0, \cdots, K-1. \end{aligned} \tag{18}$$

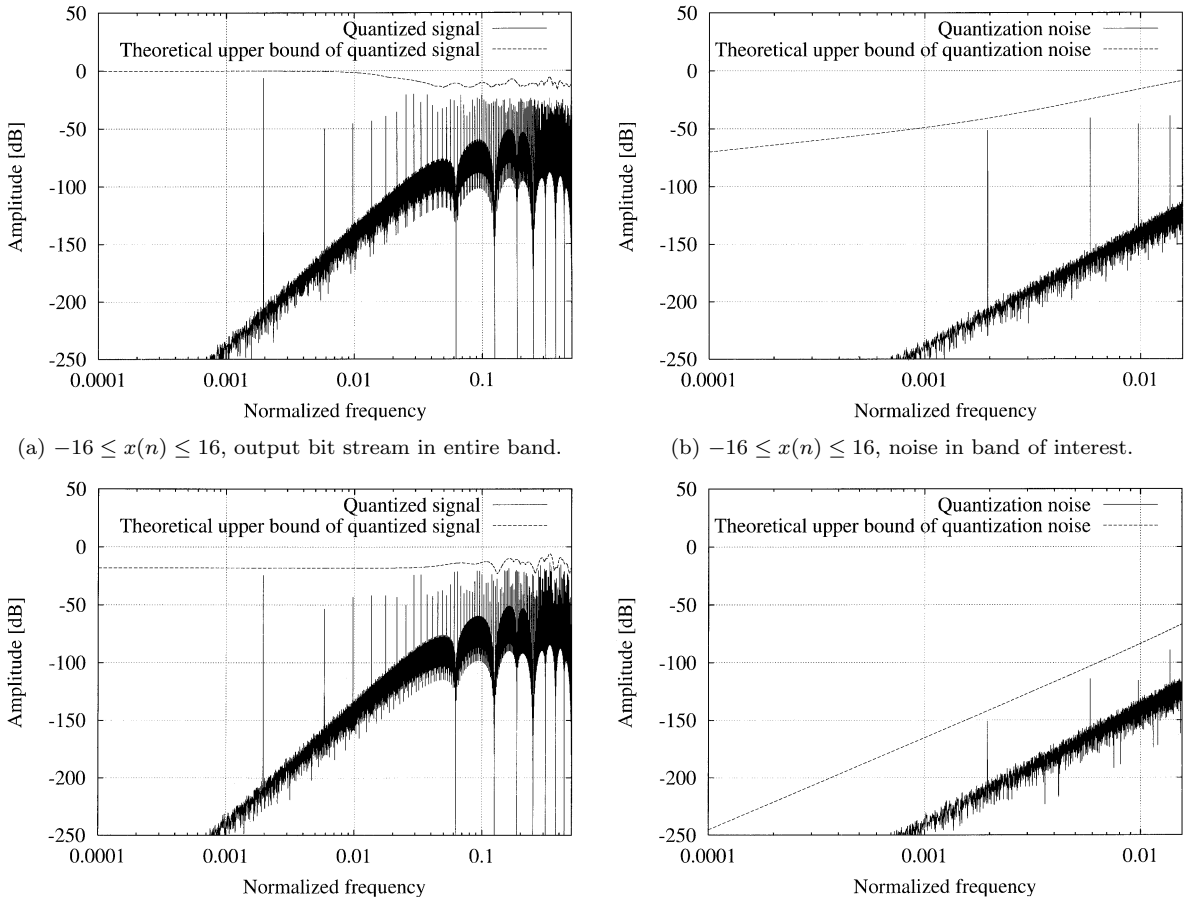
The first constraint with $k = 0$ in the optimization problem (18) guarantees that the optimized noise transfer function becomes zero at DC and has the K-th order flatness. Moreover if the objective function becomes zero with an optimum solution, the optimum noise transfer function performs $(K+1)$-th order flatness at DC. Also the constraint on the l_1 norm of $h_{NTF}(m)$ for $m = 1, \cdots, T_{NTF} - 1$ guarantees the stability of the ΣΔ modulator. As the constraint on the l_1 norm can be rewritten as a certain number of inequalities, the problem can be solved by utilizing numerical computation programs for linear programming problems [9]. With the growth of T_{NTF}, the number of those inequalities exponentially increases. So it would be hard to solve the problem with large T_{NTF} under a limited computational power. To find an appropriate upper bound c, which is dealt with as a constant in the problem, we solve the problem for a wide range of c.

4. Design Example

4.1 Single-State Binary Interpolator

First we have designed a single-state binary interpolator and then analyzed its output errors. The design specification is; $T = 32$ and its input quantization levels of -16, -15, $\cdots$, 16. The binary interpolator has been designed as a single-state machine, and all the binary vectors to be chosen for those quantization levels have a unique length l_v of 32. We have optimized the binary vectors with a wide range of τ and then determined an optimal delay of $\tau = 16$. Then using sinusoids having amplitudes of 16 or 2 with a frequency of $\pi/256$ as test input signals, we have analyzed output bit streams and quantization noise of their bit streams, and the MOES of the designed binary interpolator for two input ranges have been compared with the actual quantization noise spectra.

Figure 4 illustrates the spectra of the output bit streams and their quantization noise for the two sinusoids together with the MOES of the two input ranges $|x(n)| \le 16$ and $|x(n)| \le 2$. From Fig. 4, we find that the MOES are close to the in-band peak error and thus



(a) $-16 \leq x(n) \leq 16$, output bit stream in entire band.

(b) $-16 \leq x(n) \leq 16$, noise in band of interest.

(c) $-2 \leq x(n) \leq 2$, output bit stream in entire band.

(d) $-2 \leq x(n) \leq 2$, noise in band of interest.

Fig. 4 The quantized signals and quantization errors of the designed single-state binary interpolator.

can tightly upper-bound the actual quantization noise spectra. Also we see that the optimized binary interpolator can more accurately quantizes input signals of smaller amplitude. This is based on the fact that binary vectors optimized for inputs of small amplitude perform higher order flatness at DC in the frequency domain, since generally the error-free condition at DC for small input does not significantly reduce a set of possible binary vectors in the optimization to improve the flatness. In the results shown in Fig. 4, the single-state binary interpolator suffers from in-band tones especially with inputs of large amplitude. Input signals of small amplitude are quantized by choosing binary vectors with high order flatness at DC so that the in-band tones are significantly reduced but not perfectly removed. An in-band tone at a frequency of ω_a is derived from a periodicity of the OEES at ω_a. Since all the binary vectors of this binary interpolator have the unique delay of 16 and length of 32, with a periodicity of an input signal, such a binary interpolator periodically chooses binary vectors. Then the periodic choice of the binary vectors causes a strong periodicity of the OEES so that an in-band tone grows. A method for alleviating severe in-band tones is to successively generates a sequence of binary vectors which have many kinds of delay. Also in this method, the length of the binary vector is optimized for each input so that we can further improve the flatness of several binary vectors at DC in the frequency domain. This method will be demonstrated in the next section.

4.2 Multi-Bit ΣΔ Modulator with Multi-State Binary Interpolator

In this section, a proposed digital quantizer as shown in Fig. 1 is demonstrated. We heuristically have determined the parameters of the multi-bit ΣΔ modulator; $u_{max} = 1$, $c = 5$ and quantization levels $\{0, \pm 1, \pm 2, \pm 3\}$ and then have optimized $h_{fb}(m)$. The result is; $h_{fb}(m) = 2.091667$, -1.5875, 0.085714, 0.666667, -0.4125 and 0.155952 for $m = 1$, 3, 6, 7, 11, 13, respectively, and zero for the others. The noise transfer function with these coefficients performs the 5-th order noise shaping. Next a 6-state binary interpolator has been designed with $T = 8$. We have found a good combination of state transitions, which is de-

current state	input	next input	lv	τ	order of flatness of optimized binary vector at DC	next state
0	(0)	(*)	0	0	∞	2
0	(1)	(1,2,3)	8	7	3	0
0	(1)	(0)	16	7	4	5
0	(2)	(*)	12	7	3	1
0	(3)	(0)	20	7	3	4
0	(3)	(1,2,3)	12	7	2	1
1	(0)	(*)	0	0	∞	3
1	(1)	(0)	12	3	3	5
1	(1)	(1,2,3)	4	3	2	0
1	(2)	(0)	12	3	3	5
1	(2)	(1,2,3)	8	3	2	1
1	(3)	(0)	16	3	2	4
1	(3)	(1,2,3)	8	3	2	1
2	(0)	(*)	16	15	4	0
2	(1)	(*)	16	15	4	0
2	(2)	(*)	20	15	3	1
2	(3)	(*)	20	15	3	1
3	(0)	(*)	12	11	3	0
3	(1)	(*)	12	11	3	0
3	(2)	(*)	16	11	3	1
3	(3)	(*)	16	11	3	1
4	(*)	(*)	0	0	∞	1
5	(*)	(*)	0	0	∞	0

Fig. 5 The determined combination of state transitions. "next state" indicates the state of the binary interpolator at next time step.

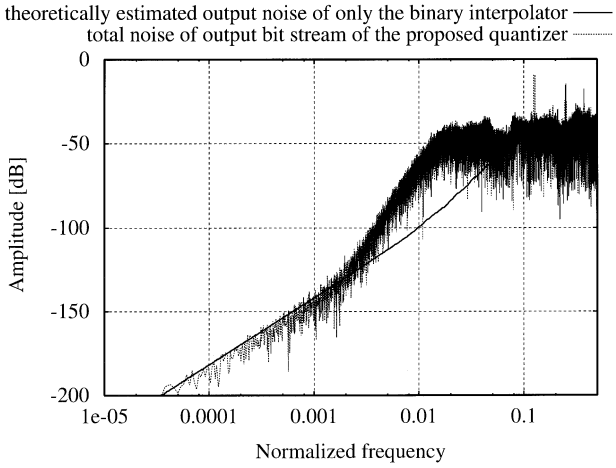


Fig. 6 Noise and distortion spectrum of the designed 3-bit ΣΔ modulator and 6-state binary interpolator with a sinusoid of an amplitude of 1.0 and a frequency $\pi/256$. The depicted spectrum is normalized so that the peak spectrum of the sinusoid can be −6.02 [dB].

picted in Fig. 5. When started up, the binary interpolator is initialized to State 1. Figure 6 illustrates the noise and distortion spectrum of an input sinusoid with the FFT of 2^{18} points. We find that the overall noise spectrum in the band of interest is shaped with two different slopes. The noise spectrum in the sharper slope region decreases by about 100 dB/decade, which should be shaped by the 5-th order ΣΔ modulator. Next we measured and estimated the probability density function (pdf) of the output of the designed ΣΔ modulator for the sinusoid. Then assuming that the output is white with the estimated pdf, we theoretically estimated the normalized $\sqrt{R_{mse}(\omega)}$ as the output noise spectrum of only the binary interpolator for the sinusoid. From Fig. 6, we see that the noise spectrum shaped by the binary interpolator is close to the theoretical estimate. Signal-to-Noise and Distortion Ratios (SNDRs) for three sinusoids which have an amplitude of 1 and different frequencies $\pi/133$, $\pi/265$ and $\pi/530$ have been evaluated. First, for the three sinusoids having three frequencies $\pi/133$, $\pi/265$ and $\pi/530$, three bands of interest have been specified as $|\omega| < \pi/128$, $|\omega| < \pi/256$ and $|\omega| < \pi/512$, respectively. Then an SNDR in each band of interest has been calculated with sharp cut-off filters in the time domain. The results have been obtained as 73.2 dB in $|\omega| < \pi/128$, 101.7 dB in $|\omega| < \pi/256$ and 120.0 dB in $|\omega| < \pi/512$. We have confirmed that the designed quantizer can achieve such high SNDRs even with the rigorous stability. Also we find that the designed 6-state binary interpolator does not generate severe in-band tones. This is derived from a fact that the 6-state binary interpolator generates binary vectors having five kinds of delay. Such a set of binary vectors having many kinds of delay corrupts a periodicity of the OEES at ω_a, which may be caused by a periodicity of an input signal to the binary interpolator. Of course by introducing many kinds of delay and length of the binary vectors, the binary vectors can also have higher order flatness at DC in the frequency domain. With these two reasons, we see that in-band tones of the 6-state binary interpolator have been significantly reduced.

5. Conclusions

In this paper, a data coding technique is first proposed, and a stable noise-shaping quantizer, which has a cascade structure of a multi-bit ΣΔ modulator and a binary interpolator, is presented. The binary interpolator chooses a pre-optimized binary vector for each input sample and successively generates the chosen binary vectors as an output bit stream. The pre-optimized binary vectors can have different lengths. The proposed binary interpolators are nonlinear recursive systems, but two methods to evaluate output errors of binary interpolators have been derived. The first method is based on the Viterbi algorithm, which enables us to upper-bound the maximum output error spectrum for all possible input signals. By using the second method, the mean squared output error spectrum can be evaluated for all possible input signals with an arbitrary probability density function. Then we have presented a method to optimize rigorously stable multi-bit ΣΔ

modulators. The optimization problem can be solved by a linear programming method.

In the design examples, first we have designed a single-state binary interpolator and analyzed its output error spectrum by the above method. Next we have designed a single-bit quantizer with a 5-th order $\Sigma\Delta$ modulator and 6-state binary interpolator. Then we have evaluated its SNDRs and a spectrum of noise and distortion. It has been confirmed that the proposed quantizer can sharply shape output noise spectra and that in-band tones can be significantly reduced with a multi-state binary interpolator. If a binary interpolator is implemented with the polyphase structure and a multiplexer, the clock speed required for the multi-bit $\Sigma\Delta$ modulator and binary interpolator can be T times slower than that of the DAC and the multiplexer, which can save the power consumption. The proposed data coding technique may be applied to [10], [11]. The papers [10], [11] have proposed quantizers with a kind of data coding technique, but they have not dealt with rigorous stability problem.

References

[1] S.R. Norsworthy, R. Schreier, and G.C. Temes, Delta-Sigma Data Converters, Theory, Design, and Simulation, IEEE Press, New York, U.S.A., 1997.
[2] I. Galton, "Spectral shaping of circuit errors in digital-to-analog converters," IEEE Trans. Circuits Syst. II, vol.44, pp.808–817, Aug. 1993.
[3] A. Yasuda, H. Tanimoto, and T. Iida, "A third-order Δ-Σ modulator using second-order noise-shaping dynamic element matching," IEEE J. Solid-State Circuits, vol.33, pp.1879–1886, Dec. 1998.
[4] S. Hein and A. Zakhor, Sigma-delta modulators, nonlinear decoding algorithms and stability analysis, Kluwer Academic Publishers, Massachusetts, U.S.A., 1993.
[5] R. Schreier and Y. Yang, "Stability tests for single-bit sigma-delta modulators with second-order FIR noise transfer functions," Proc. IEEE Int. Symp. Circuits Sys., vol.3, pp.1316–1319, May 1992.
[6] M. Yagyu, A. Nishihara, and N. Fujii, "Analysis and minimization of output errors of 2-d non-separable FIR digital filters with finite precision internal signals," IEICE Trans. Fundamentals, vol.E80-A, no.8, pp.1391–1402, Aug. 1997.
[7] G.D. Forney, Jr., "The Viterbi algorithm," Proc. IEEE, vol.61, no.3, pp.268–278, March 1973.
[8] C.W. Therrien, Discrete Random Signals and Statistical Signal Processing, Prentice Hall, NJ, 1992.
[9] R. Fletcher, Practical methods of optimization, Second Ed., John Wiley & Sons, New York, 1987.
[10] D. Birru, "Optimized reduced sample rate sigma-delta modulation," IEEE Trans. Circuits Syst. II, vol.44, no.11, pp.896–906, Nov. 1997.
[11] E. Roza, "Recursive bitstream conversion: The reverse mode," IEEE Trans. Circuits Syst. II, vol.41, no.5, pp.329–336, May 1994.

Mitsuhiko Yagyu received the B.E., M.E. and Dr.Eng. degrees in electronics from Tokyo Institute of Technology, Tokyo, Japan, in 1993, 1995 and 1998, respectively. From 1998 to 2001, he worked for Texas Instruments Japan and is now Research Associate of the Department of Physical Electronics, Tokyo Institute of Technology. His main research interests are in digital signal processing. He received IEICE Best Paper Award in 1999.

Akinori Nishihara received the B.E., M.E. and Dr.Eng. degrees in electronics from Tokyo Institute of Technology in 1973, 1975 and 1978, respectively. Since 1978 he has been with Tokyo Institute of Technology, where he is now Professor of the Center for Research and Development of Educational Technology. His main research interests are in one- and multi-dimensional signal processing, and its application to educational technology. He served as an Associate Editor of the IEICE Trans. Fundamentals. from 1990 to 1994, an Associate Editor of the IEEE Transactions on Circuits and Systems II from 1995 to 1997, and Editor-in-Chief of the Transactions of IEICE, Part A, from 1998 to 2000. He is now Educational Activities Committee Chair of IEEE Region 10 (Asia Pacific Region). He received IEICE Best Paper Award in 1999, and IEEE Third Millennium Medal in 2000. Dr. Nishihara is a member of IEEE, EURASIP, ECS and JET.