BETA EXPANSIONS: A NEW APPROACH TO DIGITALLY CORRECTED A/D CONVERSION

I. Daubechies
Department of Mathematics and
Program in Applied and Computational
Mathematics, Princeton University
Princeton, NJ 08544

R. DeVore
Department of Mathematics
University of South Carolina
Columbia, SC 29208

C.S. Güntürk
Department of Mathematics
Courant Institute of
Mathematical Sciences
New York, NY 10021

V. A. Vaishampayan
Information Sciences Research
AT&T Labs-Research
Florham Park, NJ 07932

ABSTRACT

We introduce a new architecture for pipelined (and also algorithmic) A/D converters that give exponentially accurate conversion using inaccurate comparators. An error analysis of a sigma-delta converter with an imperfect comparator and a constant input reveals a self-correction property that is not inherited by the successive refinement quantization algorithm that underlies both pipelined multistage A/D converters as well as algorithmic A/D converters. Motivated by this example, we introduce a new A/D converter—the **Beta Converter**—which has the same self-correction property as a sigma-delta converter but which exhibits higher order (exponential) accuracy with respect to the bit rate as compared to a sigma-delta converter, which exhibits only polynomial accuracy.

1. INTRODUCTION

The use of redundancy and a digital correction technique for reducing the sensitivity of A/D converters to component nonidealities (including comparator offset) was first proposed in [1]. In this approach, each stage of a multi-stage converter generates n bits, of which $(n-1)$ bits contribute to the overall resolution of the converter, whereas 1 bit is used to digitally correct for errors due to component variations. In [2], a theoretical error analysis is carried out for A/D converters with an inaccurate comparator and a modified version of the design in [1], with fewer components, is presented.

In this paper we present a conversion technique that is robust to variations in comparator voltage offset. The technique is based on computing a β-expansion for a real number, by which we mean that $y \in [0,1]$ is expressed as

$$y = \sum_{i=1}^{m} b_i \beta^{-i}, \tag{1.1}$$

where $1 < \beta < 2$ and $b_i \in \{0,1\}$. We will show that the resulting binary word has sufficient redundancy so that exponentially accurate quantization (in the length of the binary word) is obtained even with mismatches in the comparator offset voltage.

A one-bit quantizer is a mapping Q of the real numbers into the discrete set $\{-1,1\}$, defined by

$$Q(z) := \begin{cases} -1, & z \le 0 \\ 1, & z > 0. \end{cases} \tag{1.2}$$

The hardware circuit implementation of such a device is never perfect: transition often happens at some point different from 0. This results in an erroneous quantizer $\tilde{Q}$, which instead computes

$$\tilde{Q}(z) := \begin{cases} -1, & z \le \rho \\ 1, & z > \rho, \end{cases} \tag{1.3}$$

for a possibly unknown (small) value of ρ. One may further assume ρ to vary at each implementation of $\tilde{Q}$. We are interested in methods for converting an analog signal $x(t)$ defined on the real line $\mathbb{R}$ into a digital bitstream using such an imprecise one-bit quantizer.

Another essential ingredient of the methods we will be considering is the sampling operation, which maps a given signal $x(t)$ to a sequence of numbers $(x(n\tau))_{n\in\mathbb{Z}}$, where τ is the sampling interval. We assume that this operation is carried out precisely.

We shall work with bandlimited functions, i.e., functions whose Fourier transforms are compactly supported. Any bandlimited function can be recovered perfectly from its samples on a sufficiently close-spaced grid; this is known as the "sampling theorem". Let $\mathcal{S}(\Omega)$ denote the class of functions $x \in L^2(\mathbb{R})$ whose Fourier transforms are supported on $[-\Omega,\Omega]$. The Shannon-Whittaker formula gives a way to reconstruct a function $x \in \mathcal{S}(\pi)$ from its samples $(x(n))_{n\in\mathbb{Z}}$ taken on the integer grid:

$$x(t) = \sum_{n\in\mathbb{Z}} x(n) S(t-n), \tag{1.4}$$

where S is the sinc function

$$S(t) := \frac{\sin \pi t}{\pi t}. \tag{1.5}$$

The functions $S(\cdot - n)$, $n \in \mathbb{Z}$, form a complete orthonormal system for $\mathcal{S}(\pi)$. Clearly, the formula above can be extended through dilation to functions in $\mathcal{S}(\Omega)$ for arbitrary Ω.

In practice, one observes a function only on a finite portion $I = [a,b]$ of the real line $\mathbb{R}$. In addition, we shall consider functions of limited maximum amplitude; in other words, we consider the class $\mathcal{S}(\Omega, M, I)$ of all signals $x \in \mathcal{S}(\Omega)$ that take values in $(-M, M)$ when $t \in I$:

$$|x(t)| < M, \quad t \in I. \tag{1.6}$$

It will be sufficient in all of what follows to consider the case where $\Omega = \pi$ and $M = 1$. We denote $\mathcal{S}(\pi, 1, I)$ simply by $\mathcal{S}$.

In order to reconstruct the signal, each sample $x(n) \in (-1,1)$ in the expression (1.4) is simply replaced by a truncated version $\tilde{x}(n)$ of its binary expansion. (There is, however, a slight glitch

0-7803-7448-7/02/$17.00 ©2002 IEEE

to overcome due to the instability of the basis functions $S(\cdot - n)$, $n \in \mathbb{Z}$, which is reflected by the fact that

$$\sum_{n\in\mathbb{Z}} |S(t-n)| = \infty \qquad (1.7)$$

whenever t is not an integer. This can be easily fixed by oversampling, as is done in Section 2. Alternatively, one can consider functions x that have a slightly smaller bandlimit, $x \in \mathcal{S}(\alpha\pi)$ with $\alpha < 1$, in which case the sinc functions $S(t-n)$ in (1.4) can be replaced by appropriate $B(t-n)$, where B has faster decay and $\sum_n |B(t-n)| < \infty$.

Let the real number $y \in (-1, 1)$ have the binary expansion

$$y = b_0 \sum_{i=1}^{\infty} b_i 2^{-i}, \qquad (1.8)$$

with $b_0 = b_0(y) \in \{-1, 1\}$ and $b_i = b_i(y) \in \{0, 1\}$ for all $i \geq 1$. The sign bit b_0 is given by $b_0(y) = Q(y)$. The other bits can be computed using the one-bit quantizer described in (1.2) in the following algorithm known as Successive Approximation (SA). For each real number z, let $Q_1(z) := (Q(z-1)+1)/2$, i.e.,

$$Q_1(z) := \begin{cases} 0, & z \leq 1 \\ 1, & z > 1. \end{cases} \qquad (1.9)$$

Let $u_1 := 2b_0 y = 2|y|$; the first bit b_1 is given by $b_1 := Q_1(u_1)$. Then the remaining bits are computed recursively as follows: if u_n and b_n have been defined, we let

$$u_{i+1} := 2(u_i - b_i) \qquad (1.10)$$

and

$$b_{i+1} := Q_1(u_{i+1}). \qquad (1.11)$$

Let us now consider what will happen if we make errors in the quantization. We suppose that at each quantization step, the circuit does not compute $Q(z)$ but rather $\tilde{Q}(z)$, given by (1.3). This also leads to the definition

$$\tilde{Q}_1(z) := \begin{cases} 0, & z \leq 1+\rho \\ 1, & z > 1+\rho, \end{cases} \qquad (1.12)$$

where ρ may vary at each implementation of $\tilde{Q}_1$. We assume that $|\rho| \leq \delta$ where $\delta > 0$ is fixed. The debilitating effect of the quantization error can already be seen in the sign bit. Assume for example that $\rho > 0$. If $y \in (0, \rho]$, then the sign bit of y will be incorrect. No matter how the remaining bits are assigned the resulting error $|y - \tilde{y}|$ is at least as large as $|y|$ which can be as large as δ. By taking a function $x(t) = yS(t-k)$, with S the sinc function (1.5) (or $x(t) = yB(t-k)$, where B is the faster-decaying reconstruction function that can be used when $x \in \mathcal{S}(\alpha\pi)$, with $\alpha < 1$), this translates into the same possible error for PCM in its circuit implementation. Note that this is not just an anomaly of only the sign bit. For $y \in (1/2, 1/2 + \rho/2)$, the sign bit $b_0(y)$ will be correct, but the bit $b_1(y)$ will be wrong and no matter how the other bits are assigned the resulting error $|y - \tilde{y}|$ will be at least as large as $|y - 1/2|$ which could be as large as $\delta/2$.

In this paper, we shall look at two other schemes which do not suffer from this effect. In these schemes, it will be possible to reconstruct the signals perfectly by taking more bits from their digital representations, even if the quantizer used to derive the digital representations is imperfect. The compensation will come, as we shall see, from the redundancy of the codewords produced by these algorithms.

2. THE ERROR CORRECTION OF SIGMA-DELTA MODULATION

When the quantizer is imperfect as defined in (1.12), the accuracy has no asymptotic decay as m is increased, as shown by our earlier examples. In this section, we shall look at another class of encoders, given by Sigma-Delta ($\Sigma\Delta$) Modulation, which behave differently when quantization error is present.

We describe here only the simplest case of first order $\Sigma\Delta$ modulation. Let $[a, b]$ be an interval over which we wish to recover the signal x. We choose to oversample, i.e., we pick $x_n := x(n/\lambda)$ with $\lambda > 1$. Our recovery formula will be based upon an interpolation function g, whose Fourier transform $\hat{g}$ is 1 on $[-\pi, \pi]$, vanishes outside of $[-\lambda\pi, \lambda\pi]$, and satisfies

$$\sum_{n\in\mathbb{Z}} G_\lambda(n) < \infty, \qquad (2.1)$$

where $G_\lambda(n) := \sup_{t\in[0,1/\lambda)} |g(t - n/\lambda)|$.

The encoding interval is denoted $\bar{I} := [a - M, b + M]$ with M chosen such that

$$\sum_{|n| \geq \lfloor \lambda M \rfloor} G_\lambda(n) < 2^{-m}. \qquad (2.2)$$

Given (x_n), the sigma-delta encoder E_λ creates a bitstream (b_n), $b_n \in \{-1, 1\}$, $n \in \lambda\bar{I}$, whose running sums are trying to match those of x_n. There is an auxiliary sequence u_n which tracks the error between these two running sums. We take $u_n = 0$ for $n < \lambda(a - M)$ and define

$$u_{n+1} := u_n + x_n - b_n, \quad n/\lambda \in \bar{I}, \qquad (2.3)$$

where

$$b_n := Q(u_n) \qquad (2.4)$$

and Q is the quantizer defined by (1.2). We see that E_λ generates one bit for each sample, which corresponds to λ bits per Nyquist interval, therefore resulting in $|\bar{I}|\lambda$ bits for the whole signal x.

To decode the $\Sigma\Delta$ bitstream, we can use the recovery formula

$$\bar{x}(t) := \frac{1}{\lambda} \sum_{n\in\lambda\bar{I}} \bar{x}_n g(t - n/\lambda), \qquad (2.5)$$

with each $\bar{x}_n$ replaced by b_n. This gives

$$\bar{x}(t) := D((b_n))(t) := \frac{1}{\lambda} \sum_{n\in\lambda\bar{I}} b_n g(t - n/\lambda). \qquad (2.6)$$

One can easily show (see also [3]) that using appropriate g yields

$$|x(t) - \bar{x}(t)| \leq C_0 \lambda^{-1}, \quad t \in I, \qquad (2.7)$$

with the constant C_0 depending only on the choice of g. Hence we see that when the average bit rate is λ, first order $\Sigma\Delta$ modulation results in $O(\lambda^{-1})$ accuracy, much worse than what would be achieved if samples taken at the Nyquist rate were to be quantized directly. However, the remarkable fact is that $\Sigma\Delta$ encoders are impervious to error in the circuit implementation of the quantization. Theorem 2.1, below, shows that given any $\delta > 0$, then an error of at most δ in each implementation of quantization in the $\Sigma\Delta$ encoder will not affect the distortion bound (2.7) save for the constant C.

Let us see how this works. Suppose that in place of the quantizer Q of (1.2), we use the imprecise quantizer $\tilde{Q}$ of (1.3), where ρ can vary at each occurence. We assume the uniform bound $|\rho| \leq \delta$. Using these quantizers will result in a different bitstream than would be produced by using Q. In place of the auxiliary sequence u_n of (2.3) which would be the result of exact quantization, we obtain the sequence $\tilde{u}_n$, which satisfies $\tilde{u}_n = 0$ for $n/\lambda < a - M$ and

$$\tilde{u}_{n+1} = \tilde{u}_n + x_n - \tilde{b}_n, \quad n \in \lambda \bar{I} \tag{2.8}$$

where

$$\tilde{b}_n = \tilde{Q}(\tilde{u}_n). \tag{2.9}$$

We then have:

Theorem 2.1 *Suppose that $\Sigma\Delta$ modulation is implemented by using, at each occurence, one of the quantizers $\tilde{Q}$, with $|\rho| \leq \delta$, in place of Q. If the sequence $(\tilde{b}_n)$ is used in place of b_n in the decoder* (2.6), *the result is a function $\tilde{x}$ which satisifies*

$$|x(t) - \tilde{x}(t)| \leq C\lambda^{-1}, \quad t \in I, \tag{2.10}$$

with $C = C_0(2 + \delta)$ and C_0 the constant in (2.7).

Proof: This theorem was proved in [3]. We do not repeat its simple proof in detail. The main idea which is to establish the following bound:

$$|\tilde{u}_n| \leq 2 + \delta, \tag{2.11}$$

which can be proved by induction on n. It is clear for $n < \lambda(a - M)$. Assume that (2.11) has been shown for $n \leq N$. If $\tilde{u}_N \leq \rho$, then $\tilde{b}_N = -1$ and from (2.8) we have

$$\tilde{u}_{N+1} = \tilde{u}_N + x_N - \tilde{b}_N. \tag{2.12}$$

Now, $x_N - \tilde{b}_N \in [0, 2]$ and hence $\tilde{u}_{N+1} \in [-2-\delta, 2+\delta]$. A similar argument applies if $\tilde{u}_N > \rho$ and therefore we have advanced the induction hypothesis and thus proved (2.11). The remainder of the proof uses summation by parts to obtain (2.10) (see [3]). ■

The error correction capability in $\Sigma\Delta$ is related to the large amount of redundancy in the representation. The question arises whether one could utilize redundancy to build other encoders which have the best of both worlds: self correction for quantization error and exponential accuracy in terms of the bit rate. In the next section, we shall construct a class of encoders which have the flavor of PCM but rather than using the binary representation of a real number y (which is unique), they utilize representations with respect to a base $\beta \in (1, 2)$. Such beta-representations are not unique, even when β is kept fixed, and this fact is exploited to achieve the above mentioned properties.

3. BETA-ENCODERS WITH ERROR CORRECTION

We shall now show that it is possible to obtain exponential bit rate performance while retaining quantization error correction by using what we shall call *beta-encoders*. The essential idea is to replace the binary representation of a real number y by a redundant representation. A block diagram of the beta-encoder is shown in Fig. 4.1.

Let $1 < \beta < 2$ and $\gamma := 1/\beta$. Then each $y \in [0, 1]$ has a representation

$$y = \sum_{i=1}^{\infty} b_i \gamma^i \tag{3.1}$$

with

$$b_i \in \{0, 1\}. \tag{3.2}$$

In fact there are many such representations. The main observation that we shall utilize below is that no matter what bits b_i, $i = 1, \ldots, m$, have been assigned, then, as long as

$$y - \frac{\gamma^{m+1}}{1-\gamma} \leq \sum_{i=1}^{m} b_i \gamma^i \leq y, \tag{3.3}$$

there is a bit assignment $(b_k)_{k>m}$, which, when used with the previously assigned bits, will exactly recover y.

We shall use this observation in an analogous fashion to Successive Approximation to encode real numbers, with the added feature of quantization error correction. These encoders have a certain offset parameter μ whose purpose is to make sure that even when there is an imprecise implementation of the encoder, the bits assigned will satisfy (3.3); as shown below, introducing μ corresponds to carrying out the decision to set a bit to 1 only when the input is well past its minimum threshold. We let Q_1 be the quantizer of (1.9).

The beta-encoder with offset μ. *Let $\mu > 0$ and $1 < \beta < 2$. For $y \in [0, 1]$, we define $u_1 := \beta y$ and $b_1 := Q_1(u_1 - \mu)$. In general, if u_i and b_i have been defined, we let*

$$u_{i+1} := \beta(u_i - b_i), \quad b_{i+1} := Q_1(u_{i+1} - \mu). \tag{3.4}$$

It then follows that

$$\begin{aligned} y - \sum_{i=1}^{m} b_i \gamma^i &= y - \sum_{i=1}^{m} \gamma^i (u_i - \gamma u_{i+1}) \\ &= y - \gamma u_1 + \gamma^{m+1} u_{m+1} \\ &\leq \gamma^{m+1} \|u\|_{l^\infty}, \end{aligned} \tag{3.5}$$

showing that we have exponential precision in our reconstruction, provided the $|u_i|$ are uniformly bounded. We shall see below that we do indeed have such a uniform bound. Let's analyze the error correcting abilities of these encoders when the quantization is imprecise. Suppose that in place of the quantizer Q_1, we use at each iteration in the beta-encoder the imprecise quantizer $\tilde{Q}_1$ defined by (1.12) where at each application the value of ρ may vary. We assume a uniform bound $|\rho| \leq \delta$ for the quantizer errors. In place of the bits $b_i(y)$, we shall obtain inaccurate bits $\tilde{b}_i(y)$ which are defined recursively by $\tilde{u}_1 := \beta y$, $\tilde{b}_1 := \tilde{Q}_1(\tilde{u}_1 - \mu)$ and more generally,

$$\tilde{u}_{i+1} := \beta(\tilde{u}_i - \tilde{b}_i), \quad \tilde{b}_{i+1} := \tilde{Q}_1(\tilde{u}_{i+1} - \mu). \tag{3.6}$$

Theorem 3.1 *Let $\delta > 0$ and $y \in [0, 1)$. Suppose that in the beta-encoding of y, the quantizer $\tilde{Q}_1$ is used in place of Q_1 at each occurence, with the values of ρ possibly varying but always satisfying $|\rho| \leq \delta$. If $\mu \geq \delta$ and β satisfies*

$$1 < \beta \leq \frac{2 + \mu + \delta}{1 + \mu + \delta}, \tag{3.7}$$

then for each $m \geq 1$, $\tilde{y}_m := \sum_{k=1}^{m} \tilde{b}_k \gamma^k$ satisfies

$$|y - \tilde{y}_m| \leq C\gamma^m, \quad m = 1, 2, \ldots, \tag{3.8}$$

with $C = 1 + \mu + \delta$.

Proof: We first claim that

$$0 \le \tilde{u}_n \le \beta(1+\mu+\delta), \quad n = 1, 2, \ldots. \tag{3.9}$$

This is proved by induction on n. For $n = 1$ it is true because

$$\tilde{u}_1 := \beta y \le \beta.$$

Assume that (3.9) has been proved for $n = N$. If $\tilde{b}_N = 0$, then $\tilde{u}_N \le 1 + \mu + \delta$ and hence

$$0 \le \tilde{u}_{N+1} = \beta\tilde{u}_N \le \beta(1+\mu+\delta), \tag{3.10}$$

as desired. If $\tilde{b}_N = 1$, then $\tilde{u}_N > 1 + \rho + \mu \ge 1 - \delta + \mu \ge 1$. Also, in this case,

$$\begin{aligned} 0 &\le \tilde{u}_{N+1} \\ &= \beta(\tilde{u}_N - 1) \\ &\le \beta[\beta(1+\mu+\delta) - 1] \\ &\le \beta(2+\mu+\delta-1) \\ &= \beta(1+\mu+\delta) \end{aligned} \tag{3.11}$$

where we have used (3.7). This advances the induction hypothesis and proves (3.9). On the other hand,

$$\begin{aligned} y - \tilde{y}_m &= y - \sum_{k=1}^{m} \tilde{b}_k \gamma^k \\ &= \gamma\tilde{u}_1 - \sum_{k=1}^{m} \gamma^k(\tilde{u}_k - \gamma\tilde{u}_{k+1}) \\ &= \gamma^{m+1}\tilde{u}_{m+1}, \end{aligned} \tag{3.12}$$

which together with (3.9) gives (3.8). ■

(Note that, in the special case $\delta = 0$, the bound (3.9) shows that the $|u_n|$ in (3.5) are uniformly bounded, as claimed above.)

For signal encoding, we utilize the beta encoder as in PCM. Namely, we take $\lambda > 1$ and let $x_n := x(n/\lambda)$ as before. We would like to avoid keeping sign bits of the x_n. We can do this by replacing x_n by $x'_n := (x_n + 1)/2$. For each x'_n we keep the first m bits $b_1(x'_n), \ldots, b_m(x'_n)$ of the beta encoder applied to x'_n. To decode, we use the beta-encoder bits to approximately recover x'_n by

$$\tilde{x}'_n := \sum_{k=1}^{m} b_k(x_n)\gamma^k \tag{3.13}$$

and then approximately recover x_n by

$$\tilde{x}_n := 2\tilde{x}'_n - 1 \tag{3.14}$$

which satisfies

$$|x_n - \tilde{x}_n| \le C\gamma^m, \quad n \in \mathbb{Z}, \tag{3.15}$$

Given a signal x, an integer $m > 0$, and the interval I, we define $\tilde{x}$ as in (2.5) except that we use the $\tilde{x}_n$ of (3.14).

We close with the following result, which is proved in [4].

Theorem 3.2 *For any $x \in S$, the beta-encoder/decoder with m bits per sample satisfies*

$$|x(t) - \tilde{x}(t)| \le C\gamma^m, \quad t \in I, \tag{3.16}$$

with C depending only on the reconstruction filter g. Moreover, if in place of the exact quantizer Q_1, we use, at each iteration, a quantizer $\tilde{Q}_1$ given by (1.12)*, with ρ satisfiying $|\rho| \le \delta$, then we still obtain the error bound* (3.16) *with the constant C now depending also on δ.*

4. SUMMARY

We have introduced a new architecture for an A/D converter that gives exponentially accurate conversion using inaccurate comparators. The basic idea is that of using a redundant β-expansion, $1 < \beta < 2$. It is proved that the resulting redundancy is useful for recovering from offset errors in the comparator of the A/D converter.

Acknowledgment. This work was the result of a collaboration started during the Summer of 1999. The non-AT&T authors would like to thank AT&T for its hospitality and partial support (RDV and CSG). We would also like to thank the many AT&T researchers, in particular Jont Allen, with whom we had extensive and enjoyable discussions.

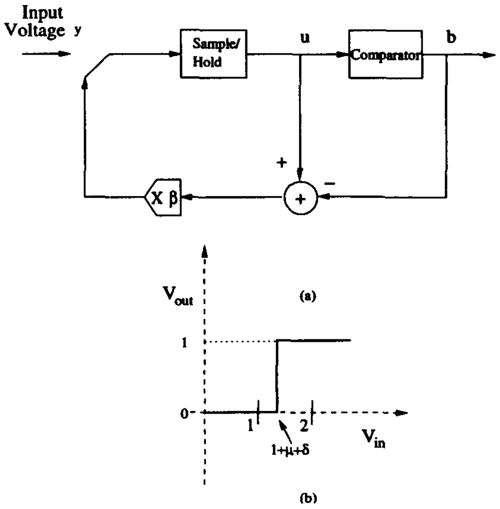

Fig. 4.1. Algorithmic β converter (a) and comparator characteristic (b).

5. REFERENCES

[1] S. H. Lewis and P. R. Gray, A pipelined 5-Msample/s 9-bit analog-to-digital converter: IEEE J. Solid-State Circuits, vol. SC-22, pp. 954-961, Dec. 1987.

[2] K. Hadidi and G. C. Temes, Error analysis in pipeline A/D converters and its applications: IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 39, No. 8 , pp. 506 -515, Aug. 1992.

[3] I. Daubechies and R. DeVore, Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order, submitted.

[4] I. Daubechies, R. DeVore, C. S. Güntürk and V. A. Vaishampayan, Exponential precision in A/D conversion with an imprefect quantizer, submitted.