Two-Step Quantization in Multibit $\Delta\Sigma$ Modulators

Saska Lindfors and Kari A. I. Halonen

***Abstract*—An architecture to simplify the circuit implementation of the internal analog-to-digital (A/D) converter in a $\Delta\Sigma$ modulator is proposed. The architecture is based on dividing the A/D conversion into two time steps, which makes the internal quantization feasible with much higher resolution than with conventional solutions. Furthermore, the time steps are interleaved so that the resolution improvement is achieved without sacrificing the speed. It is shown, with a linearized model, that the order of the noise shaping is increased by one with respect to the coarse quantization error made during the first step.**

For a high oversampling ratio, the coarse quantization error made in the first step is easily suppressed to an insignificant level due to the one order higher noise shaping. Depending on the partitioning of the bits between the conversion steps, the coarse error will dominate below a certain oversamping ratio. However, it is shown that the technique can be extended to more than one order higher noise shaping, making it useful for low oversampling ratios as well.

***Index Terms*—Analog–digital conversion, integrated circuits, quantization, sigma–delta modulation.**

I. Introduction

ONE-BIT quantization has dominated in $\Delta\Sigma$ modulators due to its inherent linearity. The linearity of the internal digital-to-analog converter (DAC) is particularly important because the overall performance of the $\Delta\Sigma$ modulator cannot be better than that. If the D/A conversion utilizes only two levels, there cannot be any mismatch between the quantization steps, and the conversion is inherently linear [1]. The circuit implementation also becomes very simple. The internal analog-to-digital (A/D) converter can be implemented with a single comparator, and the D/A converter consists of a reference voltage, a capacitor, and a couple of switches. The main drawback of using such a low resolution is the high quantization noise power generated. The signal has to be heavily oversampled in order to sufficiently suppress the quantization noise over the signal band. The noise shaping can also be made more efficient by increasing the order of the loop filter. Unfortunately, this potentially leads to problems with stability.

Despite the undisputable benefits that 1-bit quantization offers, the use of multibit (Fig. 1) quantization has recently gained widespread interest, due to the introduction of efficient dynamic element matching (DEM) techniques [2], [3]. The basic principle of these techniques is to average out the mismatch in the DAC by alternating the elements that are used for the conversion. The use of multibit quantization is desirable because the signal-to-noise ratio (SNR) can be improved without clocking the circuit very fast or getting mixed with potentially unstable systems. Furthermore, the stability problems associated with higher order $\Delta\Sigma$ modulators are alleviated by the use of multibit quantization [4]. This is because the loop gain then remains constant independent of the signal, assuming that the quantizer is not overloaded during the operation.

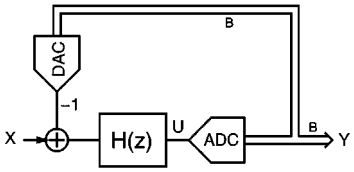

Fig. 1. A $\Delta\Sigma$ modulator with multibit internal quantization.

Low power consumption is important in all mobile electronics, but due to the limited battery size, cellular phones place particularly stringent demands on the circuit design. In $\Delta\Sigma$ modulators, the power consumption is typically dominated by the first integrator, which has to settle very accurately. The feedback signal, being high-pass shaped, contains most of its power around the Nyquist frequency, which results in tough slew-rate requirements on the amplifier. Multibit quantization relaxes the maximum current needed to integrate the feedback signal, because the height of an individual step is smaller. Assuming that the amplifiers are operating in class A, which is typically the case, the power consumption is directly reduced. However, the power consumption of the other integrators may even be higher than with 1-bit quantization because of scaling difficulties and the extra load presented by the ADC [5].

Most of the reported multibit $\Delta\Sigma$ modulators have used a moderately low number of bits in the internal quantization, although increasing the bits would have a direct impact on the overall dynamic range. The resolution is typically from 3 to 5 bits . There are two implementation-specific reasons to this.

First, the existing DEM techniques can only be applied to DACs with unweighted circuit elements. This means that the number of DAC elements and the complexity of the DEM circuitry increases exponentially with the number of bits. Most of the dynamic element matching is done by digital circuits, which are very small with the state-of-the-art silicon technologies, but the analog switching can nonetheless be complex if the number of bits is large.

Another important factor is that the A/D conversion has to be performed during a single clock cycle. The conversion result has to be available to the feedback DAC well before the next integration phase, or the loop will be unstable. This leaves the flash architecture as the only option for the internal A/D converter. Flash converters use a brute-force approach to A/D conversion, which means that the input signal

Manuscript received May 1999; revised January 2001. This paper was recommended by Associate Editor E. Soenen.

The authors are with the Electronic Circuit Design Laboratory, Helsinki University of Technology, HUT FIN-02015, Finland (e-mail: sli@ecdl.hut.fi).

Publisher Item Identifier S 1057-7130(01)03056-7.

is simultaneously compared with $2^B - 1$ reference voltages in order to decide the quantization level. Unfortunately, this means that $2^B - 1$ comparators are needed to perform the conversion (Fig. 2). Clearly, the power consumption and the area requirement of such an ADC prohibit a large number of bits. The performance of the internal ADC can be very relaxed in a $\Delta\Sigma$ modulator, because any nonlinearity is divided by the gain of the preceeding integrators. Consequently, the comparators can be much smaller and less power consuming than they would be in a stand-alone flash design. However, the last integrator in the loop filter has to be able to drive at a high clock rate the input capacitance of the flash, which can be significant. The power consumption of the last integrator can easily dominate the overall consumption if the number of bits is large. As a comparison, in a 1-bit $\Delta\Sigma$ modulator, the last integrators can be scaled much smaller than the first, and they consume only marginal power.

An arbitrarily high effective resolution can, in principle, be achieved by cascading the $\Delta\Sigma$ modulator with another $\Delta\Sigma$ modulator or an ADC [6]. However, the reduction of quantization noise relies on matching the unprecise analog loop filter of the $\Delta\Sigma$ modulator with a digital filter in the postprocessing. This is typically difficult to achieve and requires a large resolution in the $\Delta\Sigma$ modulator feedback loop to suppress the power of the quantization noise leakage.

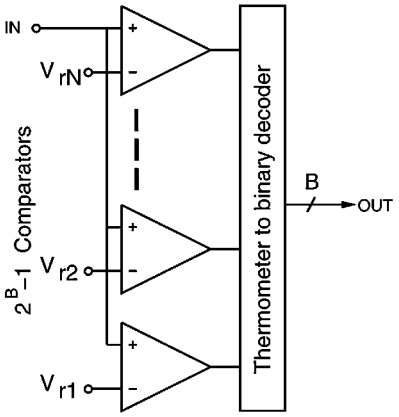

Fig. 2. A flash-type A/D converter.

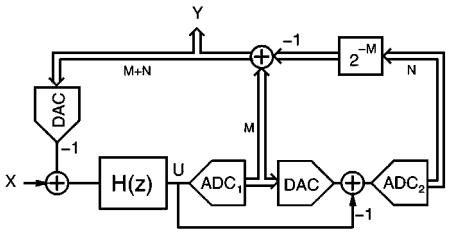

Fig. 3. A multibit $\Delta\Sigma$ modulator with the proposed two-step quantization architecture.

II. Two-Step Quantization

The $\Delta\Sigma$ modulation, as a technique, relies on oversampling, which means that all operations, like integration, A/D-, and D/A-conversion, have to be performed within roughly the same time. If any operation takes significantly longer than the others, it will limit the speed, and consequently the dynamic range.

ADCs with moderate resolution are never implemented with the flash architecture, because the requirements would become excessive. Different A/D architectures, like pipeline or successive approximation, attack this problem by dividing the conversion to two or more time steps. However, this adds latency to the signal, which cannot be tolerated inside the feedback loop of the $\Delta\Sigma$ modulator, as was pointed out in the last section.

A. The Proposed Architecture

The internal A/D conversion of a multibit $\Delta\Sigma$ modulator can be divided into two (or more) time steps if the low-resolution conversion result from the first step is fed back immediately. The incomplete A/D conversion is finalized by feeding back, during the next time step, a full-resolution corrective term representing the quantization error that remained in the coarse conversion result.

In practice, a flash converter with M-bits resolution performs the first coarse conversion (ADC_1 in Fig. 3). The output of the loop filter U is sampled by an MDAC at the same time the ADC_1 is triggered (an MDAC implements the D/A conversion and subtraction in Fig. 3). Then the difference between the coarse conversion result and the sampled loop filter output U is amplified by the MDAC, and the error is A/D converted by an N-bit flash converter ADC_2. The necessary error amplification depends on the input scales of the ADCs being 2^M for

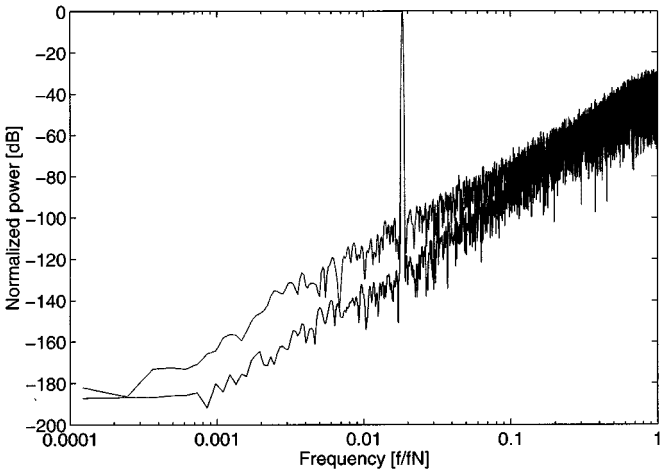

Fig. 4. The simulated spectrums for a second-order $\Delta\Sigma$ modulator with 4-bit quantization and $4 + 4$ bit two-step quantization.

equal full-scale input voltages. The two ADC stages operate in time-interleaved fashion so that the latter is always processing the loop filter output from the previous time step. The outputs from the two stages are added digitally, resulting in a feedback word of $M + N$ bits. The complexities of the ADC stages are, however, proportional only to M and N.

For example, an 8-bit flash ADC requires 255 comparators. The same resolution can be achieved with a total of only 30 comparators ($15 + 15$), which is clearly much more feasible than 255, if the 8-bit word is divided to two 4-bit subconversions. A second-order $\Delta\Sigma$ modulator with 4-bit internal quantization was simulated with MATLAB. The output code was fast Fourier transformed, and the result is shown in Fig. 4 as the upper curve.

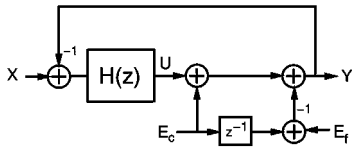

Fig. 5. The signal flow graph of the proposed architecture.

The calculated SNR ratio was 91.3 dB with -6 dBr input ampltiude and oversampling ratio of 64. The simulation result of the proposed architecture with $4+4$-bit quantization is also plotted in Fig. 4. It can be seen from the figure that the quantization noise floor is reduced at lower frequencies. The calculated SNR with the same parameters was 116.2 dB, which is 24.9 dB lower, as could be expected with four extra bits.[1]

At higher frequencies, the noise floor of the two-step quantization rises more steeply. Finally, close to the Nyquist frequency, it exceeds the noise floor of the lower resolution $\Delta\Sigma$ modulator. This effect will be investigated more fully in the following sections.

B. The Noise Transfer Functions

Let us study a linearized model of the proposed architecture, where the quantizers ADC_1 and ADC_2 in Fig. 3 are modeled as white noise sources. This model is valid if the quantization error is evenly distributed and does not correlate with the signal, which is a much more reasonable assumption with multibit quantization that with 1-bit quantization. The coarse quantization is modeled as a white noise source $E_c(z)$ having a power of

$$e_c^2 = \frac{2^{-2M}}{12} V_{\text{ref}}^2 \tag{1}$$

where it was assumed that the quantizer input range is from $-V_{\text{ref}}/2$ to $+V_{\text{ref}}/2$. The quantization error is added to the output of the loop filter U in Fig. 5 and fed to the second quantizer, which is modeled as another additive white noise source $E_f(z)$. The quantization step of ADC_2 is given by $\Delta_c/2^N$, and consequently, the power of the fine quantization error $E_f(z)$ is found to be

$$e_f^2 = \frac{2^{-2(M+N)}}{12} V_{\text{ref}}^2. \tag{2}$$

In Fig. 3, the coarse error is formed by D/A converting the digital output of ADC_1 and subtracting it from the loop filter output. The coarse error is quantized by ADC_2 during the next time step, which is reflected by the delay in Fig. 5. In Fig. 3, the output from ADC_2 is divided by 2^M and subtracted from the ADC_1 output to give the feedback word $Y(z)$. The division is not shown in Fig. 5 because it is already included in the power of the fine quantization error $E_f(z)$ of (2).

Let us write the output signal $Y(z)$ from Fig. 5

$$Y(z) = \frac{H(z)}{1+H(z)} X(z) + \frac{(1-z^{-1})E_c(z) + E_f(z)}{1+H(z)} \tag{3}$$

[1]The one extra decibel is attributed to the random nature of the $\Delta\Sigma$ modulator, which may result in small variation of the results.

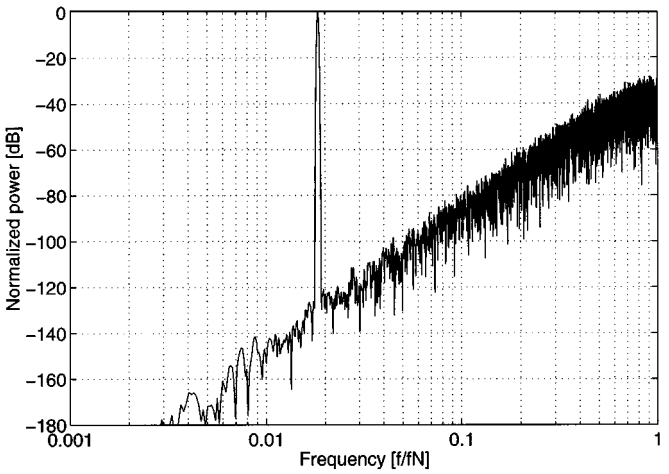

Fig. 6. The simulated spectrum with $4+\infty$ bits in the quantization showing the third-order shaped coarse quantization error.

where $X(z)$ is the input signal and $H(z)$ is the loop filter. From (3), we may note that there are two different noise transfer functions (NTFs), and that that of the coarse error is one degree higher than the loop filter due to the extra differentiation. In practice, this means that the coarse error is not completely canceled, but it is just shaped more aggressively. If the final quantization resolution $M+N$ is much higher than the resolution M of the first ADC stage, the coarse A/D conversion will limit the resolution.

The coarse error can be seen from Fig. 6, which shows the resulting spectrum when the resolution of the fine quantization is taken to infinity. As expected, the slope is 60 dB/decade. Comparing the SNR, which is 123.4 dB for an oversampling ratio (OSR) of 64 with the $4+4$-bit case, it can be seen that the fine quantization error dominates for this OSR, but the coarse error power is only 7 dB lower.

At low oversampling ratios, the coarse quantization error will take over due to the higher order of its shaping function. The noise contribution of the coarse quantization error will dominate at frequencies where

$$|\text{NTF}_c(z)|e_c > |\text{NTF}_f(z)|e_f. \tag{4}$$

From (3), the coarse error transfer function is given by

$$\text{NTF}_c(z) = (1-z^{-1})\text{NTF}_f(z). \tag{5}$$

Combining (4) and (5), we get

$$|1-z^{-1}| > \frac{e_f}{e_c}. \tag{6}$$

Assuming that the frequency range of interest is much lower than the Nyquist frequency f_N, we may approximate (6) by

$$\pi \frac{f}{f_N} > \frac{e_f}{e_c}. \tag{7}$$

Combining (1), (2), and (7), we get the crossover oversampling ratio below, which the coarse error dominates

$$\text{OSR} > \pi 2^N. \tag{8}$$

In the case of 4 + 4-bit quantization, (8) gives 50.2 as the crossover oversampling ratio.

The OSR can be traded for the implementation complexity by moving some bits to the coarse quantization step. The difference in the transfer function slopes is 6 dB/octave, which means that for each additional bit, the crossover OSR is halved. If the bits are added to the coarse quantization, so that the overall resolution is increased, the crossover OSR remains unaltered.

C. Stability

The feedback loop is stabilized by the fact that most of the signal power is fed back without latency and only a small corrective term with an extra delay. The stability of the feedback loop depends on the poles of the signal transfer function and NTFs, which vary as a function of the quantizer gain [7]. Let us define the quantizer gain as a function between the analog input u of the ADC and its digital output y

$$\lambda(k) = \frac{y(k)}{u(k)} = 1 + \frac{e(k)}{u(k)} \tag{9}$$

where $e(k)$ is the quantization error. In a $\Delta\Sigma$ modulator, the instability occurs, when the quantizer gain gets too low [8]. Normally, with a multibit quantizer, the gain remains bounded above $1+e/y_{\max}$, assuming that the quantizer does not saturate. The fine quantization error gives only some extra uncertainty to the quantizer gain, which may, as a worst case, double the variation. If the loop filter is designed so that there is margin for this extra variation, the resulting $\Delta\Sigma$ modulator will be stable. A second-order $\Delta\Sigma$ modulator with the proposed two-step quantization was simulated for input signal frequencies ranging from 2% of the Nyquist frequency up to 99%. The same simulation was performed for a signal amplitude of 6 dB below the reference (Fig. 7, solid line) and with a signal amplitude equal to the reference (dashed). No indication of instable operation due to the third-order shaping for the coarse quantization error was found, as can be noted from the reasonably flat SNR curves in Fig. 7.

D. Matching

The fact that there are two different A/D conversions, which are by nature nonprecise operations, gives rise to some error. The simplest way to model this is to assume that the gain of the fine quantization deviates from unity by a small gain error ΔA. The coarse error transfer function is then given by

$$\mathrm{NTF}_c(z) = \frac{1-(1+\Delta A)z^{-1}}{1+H(z)}. \tag{10}$$

The extra transmission zero in the coarse error transfer function is shifted away from dc, limiting the attenuation at frequencies below zero. If we require that the coarse error be suppressed K decibels below the fine quantization error, the maximum gain error is then given by

$$\Delta A < 10^{-[N\cdot 6.02\,\mathrm{dB}+K]/20}. \tag{11}$$

For the case of 4 + 4-bit quantization and 3-dB margin, the gain error must be less than 4.4%, which is not a very tough requirement. Furthermore, if the zero is much below the signal bandwidth, the fine quantization close to the upper edge of the signal will dominate and small leakage of coarse quantization error at lower frequencies is of no consequence. For the 4.4% mismatch, the zero is at $0.0134 f_N$, which corresponds to an OSR of 75.

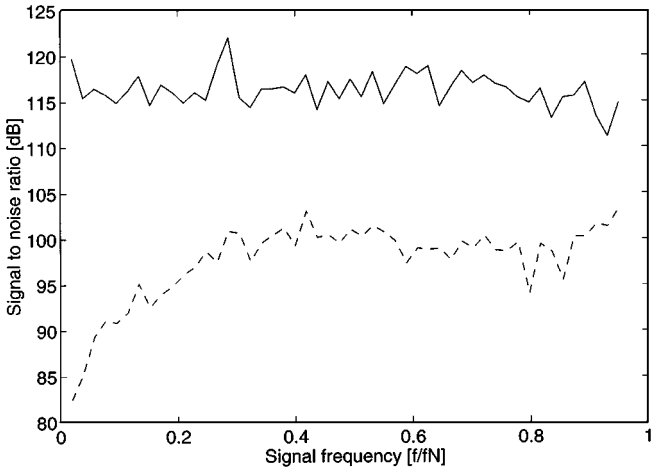

Fig. 7. Simulated SNR as a function of signal frequency (solid −6 dBr input, dashed 0 dBr input).

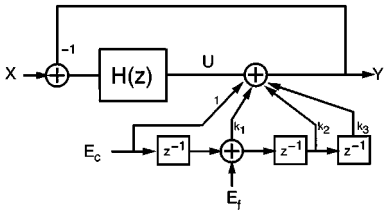

Fig. 8. The signal flow graph of the proposed architecture with a 4-tap coarse error shaping finite impulse response.

E. Higher Order Coarse Error Shaping

When a full conversion of a sample at the input of the internal quantization has been performed, the information of the coarse error can be used at a later time to further improve its shaping. Using one extra sample, we get a three-tap finite impulse response (FIR) filter, which increases the order of the coarse error transfer function $\mathrm{NTF}_c(z)$ by two instead of one. Although this works beautifully at a transfer function level, the noise accumulation due to 24-dB gain at the Nyquist frequency tends to saturate the feedback loop, rendering it useless. This problem can be solved by adding two taps to the coarse error-shaping FIR instead of one (Fig. 8). Then two transmission zeros can be placed at low frequencies to improve the SNR over the signal band and one zero to the Nyquist frequency to prevent the out-of-band quantization noise from saturating the loop.

The extra branches in the coarse error-shaping FIR can be implemented entirely in the digital domain by storing output words from ADC_2 to a shift register and calculating weighted sums, according to the FIR tap coefficients. This requires digital multiplication operations, which increase the circuit complexity. However, the multiplications are all fixed, so that canonic signed digita arithmetics may be utilized to eliminate the need for real digital multipliers.

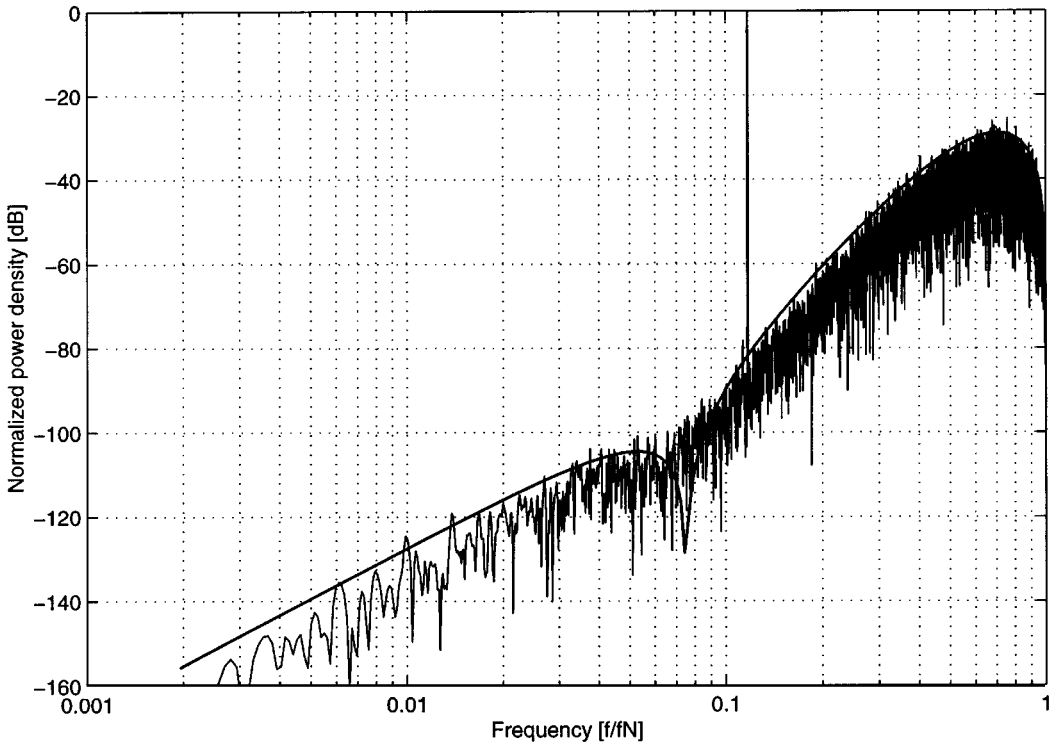

Fig. 9. The noise power density with 4-tap coarse error shaping FIR (1, −0.945, −0.945, 1). The solid line represents the coarse error transfer function.

The output spectrum of a $\Delta\Sigma$ modulator with $4+4$-bit quantization and a 4-tap coarse error shaping FIR is plotted in Fig. 9. The integrated noise power is 75.8 dB below a −6-dBr signal for an oversampling ratio of 12. The benefit of a higher order noise transfer function diminishes at low oversampling ratios if all the zeros are placed at dc. Therefore, the low-frequency zero pair is optimized so that the noise power density due to the coarse error is more flat over the signal band. The notch produced by this zero pair is not visible as it is flooded with noise from the fine quantization.

Unfortunately, the fine error transfer function $\mathrm{NTF}_f(z)$ is also changed, as it must share a part of the same FIR with the coarse error. Let us rewrite the noise transfer function for a $\Delta\Sigma$ modulator with a general coarse error-shaping FIR $L(z)$

$$Y(z) = \frac{L(z)}{1+H(z)}E_c(z) + \frac{1-L(z)}{1+H(z)}E_f(z). \qquad (12)$$

In order to minimize the coarse error power $E_c(z)$, it is desirable to design $L(z)$ to have large attenuation over the frequency range of interest. Assuming that $L(z) \ll 1$, it follows that the fine error transfer function is not changed much and that it is approximately equal to (3). As an example, the numerator $N(z)$ of the fine error transfer function becomes

$$N(z) = z^{-1}(0.945 + 0.945z^{-1} - z^{-2}) \qquad (13)$$

for the simulated four tap FIR with coefficients 1, −0.945, −0.945, and 1. For an OSR of 12, the fine quantization error power is increased by less than 0.2 dB, which can be considered marginal.

III. Conclusions

The internal resolution of the $\Delta\Sigma$ modulators has been limited by the exponential cost function associated with the number of bits. The proposed architecture enables the realization of moderate-resolution internal A/D converters, with power consumption and area that would otherwise be prohibitively high.

The one order higher shaping of the coarse quantization error suppresses it significantly at low frequencies, but attention must be paid to the partitioning of the bits between the coarse and fine quantization and the oversampling ratio in order to prevent the one order higher shaped coarse error from dominating the dynamic range. The order of the coarse error shaping can be further increased by using more taps in the coarse error-shaping FIR. This extends the usefulness of the proposed technique to low oversampling ratios.

The realization of the internal DAC has basically the same problem of being exponentially more expensive as the number of bits is increased. This problem still has to be addressed somehow. However, most of the signal processing of DEM algorithms is in the digital domain, and the analog unit elements of the DAC are basically single components (capacitors or MOS transistors), which makes them smaller than the comparators in the ADC.

References

[1] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta–Sigma Data Converters*. New York: IEEE Press, 1992.

[2] B. H. Leung and S. Sutarja, "Multibit $\Sigma-\Delta$ A/D converter incorporating a novel class of dynamic element matching techniques," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 35–51, Jan. 1993.

[3] R. T. Baird and T. S. Fiez, "Linearity enhancement of multibit $\Delta\Sigma$ A/D and D/A converters using data weighted averaging," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 753–762, Dec. 1995.
[4] J. G. Kenney and L. R. Carley, "Design of multibit noise-shaping data converters," *Int. J. Analog Integr. Circuits Signal Process.*, pp. 259–272, 1993.
[5] S. Lindfors, M. Länsirinne, T. Lindeman, and K. Halonen, "On the design of multibit $\Delta\Sigma$-modulators," in *Proc. ISCAS'99*, vol. 2, 1999, pp. 13–16.
[6] T. Brooks, D. Robertson, D. Kelly, A. Muro, and S. Harston, "A cascaded Sigma–Delta pipeline A/D converter with 1.25 MHz signal bandwidth and 89dB SNR," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1896–1906, Dec. 1997.
[7] E. Stikvoort, "Some remarks on the stability and performance of the noise shaper or Sigma–Delta modulator," *IEEE Trans. Commun.*, vol. 36, pp. 1157–1162, Oct. 1988.
[8] R. T. Baird and T. S. Fiez, "Stability analysis of higher-order Delta–Sigma modulation for ADC's," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 59–62, Jan. 1994.

Saska Lindfors received the M.Sc., Lic.Tech., and Dr.Sc. degrees in electrical engineering from the Helsinki University of Technology (HUT), Finland, in 1994, 1998, and 2000, respectively.

After receiving the M.Sc. degree, he was with the ASSP Group, Fincitec Ltd., as an Integrated Circuit Designer. During 1995–1999, he was with the Electronic Circuit Design Laboratory of HUT as a Research Engineer and Teaching Assistant. In 2000, he joined the Radio Electronics Laboratory of the Royal Institute of Technology, where he is an Assistant Professor in the area of integrated circuit design. He has published more than 30 technical papers on circuit techniques related to baseband filtering, delta–sigma modulation, and high-speed sampling. He has received one patent and has two patents pending.

Kari A. I. Halonen was born in Helsinki, Finland, on May 23, 1958. He received the M.Sc. degree from Helsinki University of Technology (HUT), Finland, in 1982 and the Ph.D. degree from the Katholieke Universiteit Leuven, Heverlee, Belgium, in 1987, both in electrical engineering.

From 1982 to 1984, he was an Assistant with HUT and a Research Assistant with the Technical Research Center of Finland. From 1984 to 1987, he was a Research Assistant with the E.S.A.T. Laboratory of the Katholieke Universiteit Leuven, enjoying also a temporary grant of the Academy of Finland. Since 1988, he has been with the Electronic Circuit Design Laboratory, HUT, as a Senior Assistant (1988–1990) and Director of the Integrated Circuit Design Unit of the Microelectronics Center (1990–1993). He was on leave of absence during academic year 1992–1993 as R&D Manager in Fincitec Inc., Finland. From 1993 to 1996, he was an Associate Professor and since 1997 has been a Full Professor in the Faculty of Electrical Engineering and Telecommunications, HUT. He specializes in CMOS and BiCMOS analog integrated circuits, particularly for telecommunication applications. He is an author or coauthor of 100 international and national conference and journal publications on analog integrated circuits.