Audio Engineering Society

Convention Express Paper 267

Presented at the AES 157th Convention
2024 October 8-10, New York, NY, USA

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Improved Analogue-to-Digital converter for High-Quality Audio

Craven, Peter[1] *AES Fellow*, Stuart, J. Robert[2] *AES Fellow*

[1] *Algol Applications, UK*

[2] *Consultant, UK*

Correspondence should be addressed to Robert Stuart (jrsmeridian@gmail.com)

ABSTRACT

A high-oversampled, low-bit modulator typical of modern audio ADCs needs a downsampler to provide PCM audio at sampling rates ranging from 44.1 kHz to 768 kHz. Traditionally, a multistage downsampler requantises at each stage, raising questions about audio transparency.

We present a decimator design in which there is no requantisation other than a single dithered quantisation when necessary to produce a final audio output of finite precision such as 24 bits. All processing is minimum-phase and concordant with the principles introduced in [1] which optimize for a specific compact impulse response and minimal (zero) modulation noise. [2]

1 Introduction

In [1] the authors described a hierarchical method for recording, archiving and efficient distribution of high-quality audio.

By considering the properties of the signal and of the receiver, we can usefully exploit modern ideas in digital sampling that lead to more precise capture of micro-temporal elements in the sound which seem to be significant for the human listener, while at the same time, encoding within a more compact data space.

Central to the theory is the way in which auditory neurons respond to the complexity of sounds of the natural world. Important classes of stimuli are the background sounds of the environment, animal vocalizations and speech. All three exhibit causality, self-similarity and a general spectral tendency for amplitude to fall with frequency; environmental sounds show a 1/f trend. Music has similar properties.

This paper continues the thread from [1], focusing on the beginning of the chain – digitizing analogue signals.

1.1 ADC Architecture

When it comes to converting these signals to digital, most modern ADCs use the high-oversampling methodology (also known as $\Sigma\Delta$ or $\Delta\Sigma$) whereby a modulator operating at e.g., ~6 MHz, ~12 MHz or ~24 MHz, provides a noise-shaped signal with a restricted wordwidth such as 4, 5, 6, 7 or 8 bits.

Figure 1 shows the generic architecture, which is popular because it requires minimal analogue pre-filtering and readily integrates into a single chip. It is an extremely difficult (analogue) problem to build a native-rate (flash) converter that maintains precise bit weighting for bit-depths such as 16b or 24b. However, integrated circuits can yield a converter with good matching up to 6 bits at a high rate. Noise-shaping enables the quantization noise and dither from this conversion to be spectrally shaped to give good dynamic range within the desired output frequency range at the output of a decimation chain.

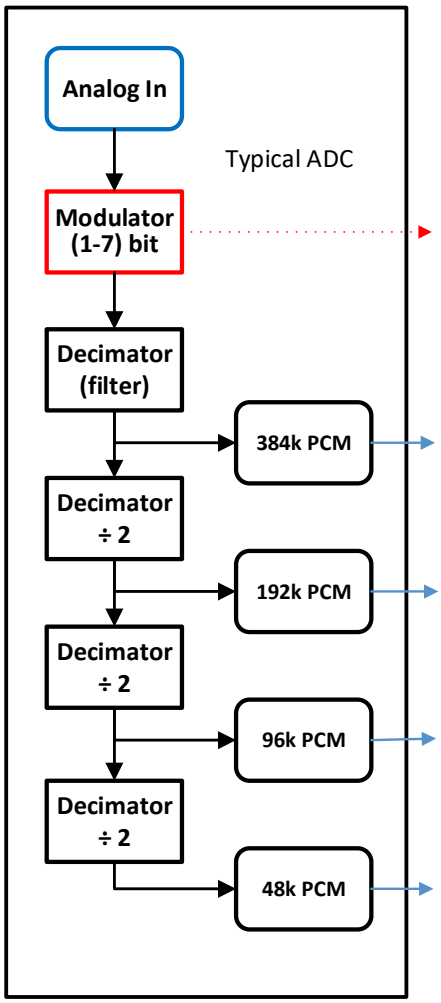

Figure 1 Typical Analog-to-Digital Converter

Although the quantiser can use fewer bits, as low as 1 bit, there are disadvantages. The fewer the bits the more noise-shaping is required, often impacting the noise floor below 20 kHz. Crucially, a 1-bit quantizer cannot be dithered and so the conversion is never linear or free of modulation noise. For a high-performance converter the minimum bitwidth is 3, with 2 used for the TPDF dither. [2]

Viewed wideband, the modulator signal is inevitably very noisy, but the noise shaping pushes the noise into the very high ultrasonic region. During decimation a lowpass filter removes the noise so the signal can then be resampled to provide high resolution at a standard audio sampling frequency between 44.1 kHz and 768 kHz.

Conversion from ~24 MHz to ~48 kHz implies a frequency ratio of 512:1 and in practice this is broken down into stages. In the limit, one could use nine stages in a decimation chain, each downsampling by 2:1; in practice, three stages may be typical. Each filter will typically produce a wide wordwidth that needs to be requantized to a smaller width acceptable to the next stage or to the final output.

Each requantisation requires dither [2], yet high-quality dither may be expensive at the most highly oversampled rates. Conversely, at rates closer to the audio band we wish to avoid the perceived 'temporal blur' associated with resampling filters and dither convergence.[2] These steep filters have extended impulse responses, normally including pre-responses which are considered perceptually unfortunate.[4]

1.2 The Decimation Chain

This paper focusses on the signal chain following the modulator and shows how a cascade of integer-coefficient moving-average filters can downsample while suppressing aliased modulator noise and minimising temporal blur. We also show how the cascade can be implemented efficiently within a CIC framework.

Here we can exploit modern sampling ideas based on Finite-Rate-of-Innovation [3] and be guided by modern neuroscience [1] which values time-domain performance at least as highly as frequency-domain performance. This approach contrasts sharply with established practice – which requires anti-alias and anti-imaging filters to have steep rates of cut-off.

2 Integer decimation with CIC filters

CIC (Comb-Integrator-Comb) architectures [6][8][9], are extensively used for the early stages of ADC decimation chains. They are seldom or never used for the later stages because of the tradition of requiring steep cut-off rates. However, they provide an economical way to construct filters having a tight time-response on account of their kernels having a close relation to B-splines rather than sinewaves. [1][7] A filter that takes the sum of two adjacent samples has z-transform response:

$$(1 + z^{-1}) \qquad (1)$$

This has gain 2 at DC so the filtered signal requires an extra headroom bit.

Assuming an initial sampling frequency of 24 MHz, the filter has zero response at 12 MHz.

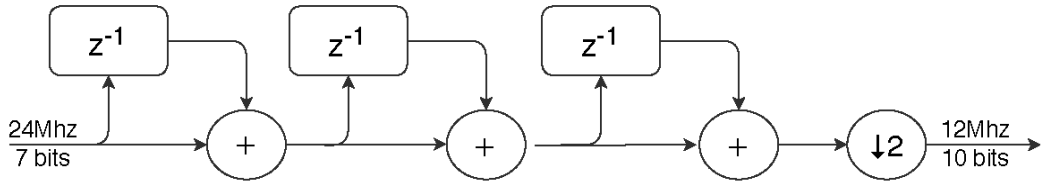

Figure 2 Downsampling-by-two using $(1 + z^{-1})$

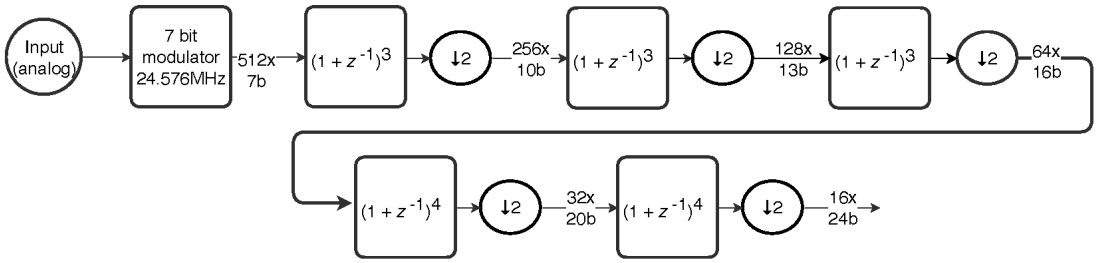

Figure 3 Unoptimized integer decimation from 512x to 16x using filter orders [3, 3, 3, 4, 4]

Three such filters in cascade, as in Figure 2, have transfer function $(1 + z^{-1})^3$ which provides sufficient attenuation in the neighborhood of 12 MHz that the filtered signal can be resampled at 12 MHz without introducing significant aliased noise into the audio band.

We have thus halved the sampling rate at the expense of increasing the bitwidth by three, so now we have 10 bits at 12 MHz if we started with 7 bits at 24 MHz.

We can repeat the same trick to bring the sampling frequency down to 6 MHz, so the bitwidth is now 11 + 3=14. And again, to produce a 3 MHz-sampled signal requiring: 14 + 3=17 bits.

And so on, except that at the lower rates the $(1 + z^{-1})^3$ filter may not provide enough attenuation to control aliasing and we may prefer $(1 + z^{-1})^4$.

Figure 3 shows a pedantic implementation of this concept, taking a signal from a 24.576 MHz modulator ('512x') and delivering 24-bit audio at 768 kHz ('16x'). [1]

3 CIC Implementation

Figure 4 (a) is a CIC implementation of the first three stages of Figure 3, reducing the rate from 512x to 64x. The components 'I' (digital integrator) and 'D' (digital differentiator) are also shown, having responses $1 / (1 - z^{-1})$ and $(1 - z^{-1})$ respectively.

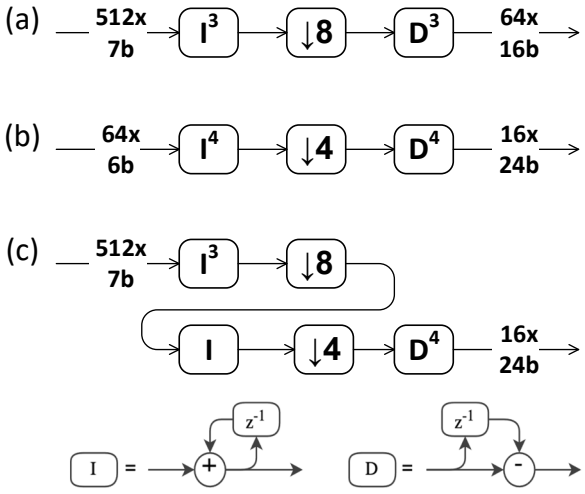

Figure 4 CIC decimation (a) from 512x to 64x and (b) from 64x to 16x (c) from 512x to 16x.

[1] Throughout this paper we refer to sampling frequencies as multiples of the baseband (typically 44.1 or 48 kHz). Commonly the intermediate division steps in an ADC will be integer multiples, up to and including the modulator itself (512x 48 kHz in this example).

'I^3' is a cascade of three 'I's and 'D^3' is a cascade of three 'D's. ' ↓ 8' is decimation by 8.

In each case 'z' represents a one-sample delay *at the local sampling rate* so D^3 (at 64x) does not cancel I^3 (at 512x) but combines with it to produce a spline-like total response.

Figure 4 (b) similarly implements rate reduction from 64x to 16x but with (conceptually) a transfer function $(1 + z^{-1})^4$ at each stage.

One might be concerned that 'I^3' has infinite gain at DC so its output would surely overload any subsequent processing. However, the magic of CIC filtering [9] is that if all the signal paths use 'wrap-round' (not saturating) arithmetic), with headroom sufficient to accommodate the final decimated output, then the correct value will be recovered. (Mathematically, we are using modulo arithmetic.)

To reduce the sample rate from 512x to 16x we could simply concatenate Figure 4 (a) and (b), but this would result in a 'D^3' followed immediately by an 'I^4'. These two operations can be elided into a single 'I', as shown in Figure 4 (c).

The only arithmetic operations required by these designs are addition and subtraction, facilitating efficient implementation on custom silicon. On a general-purpose processor it may be preferred to implement as standard FIR filters, for example expanding $(1 + z^{-1})^3$ as $(1 + 3z^{-1} + 3z^{-2} + z^{-3})$.

To describe decimation filtering of Figure 3, we use the notation:

[3, 3, 3, 4, 4]

as a list, for each intermediate sampling frequency, starting with the modulator frequency, the power to which a $(1 + z^{-1})$ element is raised prior to subsampling down to the next frequency.

The notation describes the response, not the implementation, and will also be used for CIC filters.

It will be apparent that the increase in bitwidth from decimation equals the sum of the integers in the list, i.e. 17 bits in the above case.

4 Aliased noise at 16x

The 512x → 16x decimation chain introduces no noise of its own but we need to assess how much modulator noise will be aliased downwards.

The unshaped noise power from a 7-bit quantiser with TPDF dither is −39 dB wideband relative to full-scale sinewave, or –37 dBFS if we cautiously allow 2 dB headroom for noise-shaped quantiser excursion. If clocked at 512x (24.576 MHz) the noise within the audio band 0–20 kHz would be −65 dBFS.

We consider a 4th-order maximally flat (Butterworth) shaper with all its zeroes at the origin and a –3 dB knee at 1 MHz. The resulting noise transfer function, shown later as the black trace in Figure 9, plateaus to +2.9 dB above 1 MHz. At lower frequencies it follows a 24 dB/8ve slope and is −133 dB at 20 kHz. So if there were no aliasing the modulator noise within the audio band would be less than: (−65 –133) = −198 dB, and actually −205 dB integrated over 0–20 kHz.

To assess the noise, we note that alias noise is introduced at every '↓2' downsampling stage in Figure 3, or at the '↓4' and '↓8' downsamplings in Figure 4. Fortunately, some further mathematical magic reveals that the decimation chain is equivalent to single linear filter at 512x followed by selection of every 32nd sample to bring the rate to 16x, so that final selection is all that needs to be considered.

The noise spectrum at the filter output is shown in Figure 5, with the parts that will alias into the audio range 0−20 kHz highlighted in red. There is a zero at DC because of the noise shaping and a zero at 768 kHz and all its multiples because of the CIC filter. These zeroes control the noise aliased into the audio range when subsampling to 16x.

The resulting noise spectra within the subsampled 16x signal are shown in Figure 6. Noise densities are expressed as dB in a 20 kHz bandwidth so are directly comparable with a specification for analogue noise such as '−130 dB in 0–20 kHz', shown dotted. It will be seen that filtered modulator noise including alias (blue) does not reach this density until 126 kHz.

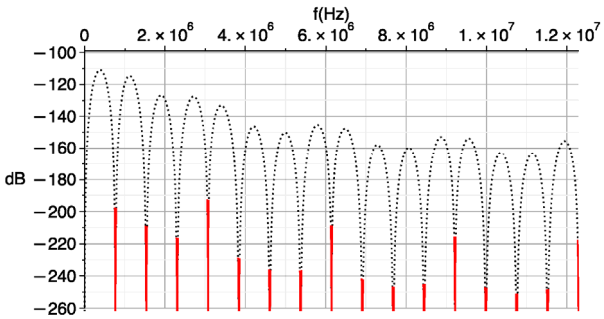

Figure 5 Model of filtered modulator noise

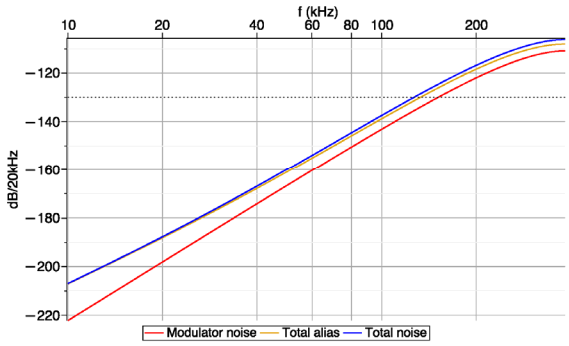

Figure 6 Aliasing @ 16x

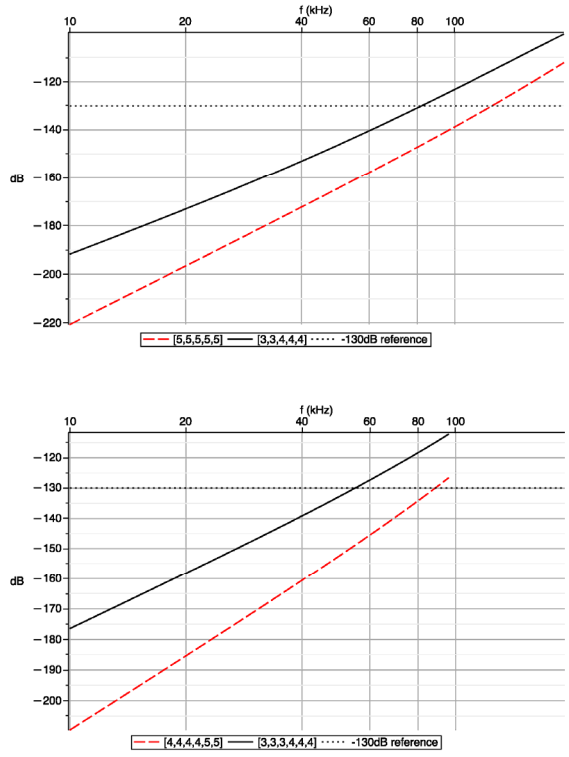

Figure 7 Spectra of alias contributions for configurations from Appendix 1, audio outputs: (upper) 384 kHz, and (lower) 192 kHz.[2]

5 Quantisation-free output at 16x

The 24-bit 16x signal downsampled using the [3, 3, 3, 4, 4] recipe can be exported directly: an audio workstation can thus receive a signal that is completely unquantized (save for the quantization inherent in the modulator).

It has a frequency response of –0.04 dB at 20 kHz: this can be corrected if desired within the workstation using 64-bit floating point arithmetic.

We can now see why the [3, 3, 3, 4, 4] recipe was proposed. [4, 4, 4, 4, 4] would give marginally greater alias rejection and allow implementation as a single CIC instead of the two downsampling stages shown in Figure 4 (c). However, the unquantized signal would require 26 bits so to export and would need a more advanced output interface, such as one capable of delivering 32-bit integer data.

6 Output at all sample rates

Subject to the capability to handle longer word widths, CIC downsampling can be continued to 8x, 4x, 2x and 1x but several questions arise:

1) How does CIC filtering sound when used to resample down to frequencies closer to the audio band?
2) Is the attenuation sufficient to keep ultrasonic interference (e.g. from lighting or radio) out of the audio band?
3) Can we still export without requantisation?
4) Can we export a flattened (i.e. droop-corrected) signal with just one requantisation?

6.1 Sonic Quality

Regarding the first question, the authors assert, on the basis of extensive listening to audio downsampled to 2x or to 1x using low-order minimum-phase filters, that the sound quality can be excellent, and certainly superior to that of ‘brick-wall’ filtering. [4] This can be explained in terms of close relationship between CIC filtering and the spline-based filtering in reference [1], especially §4.3 and §4.4 therein.

6.2 Sensitivity to Ultrasonics

Figure 8 addresses the most demanding case of downsampling to 1x, showing two CIC frequency responses, one ‘minimal’ (red) recipe configured for minimal ‘blur’ and the other ‘safe’ giving greater ultrasonic rejection.

As the ear’s sensitivity falls rapidly above 15 kHz, the most vulnerable situation is probably of a 33 kHz tone aliasing to 15 kHz. The ratio of the CIC responses at these two frequencies indicates the alias suppression: it is 20 dB for the ‘minimal’ recipe shown red in Figure 8 and 45 dB for the ‘safe’ recipe.

[2] As mentioned, for clarity the noise density plots in Figure 6 and Figure 7 are offset by 43 dB so that, rather than show noise-density in per-root-Hz, the numbers indicate the level that a white-spectrum noise of that density would total in a 20 kHz bandwidth.

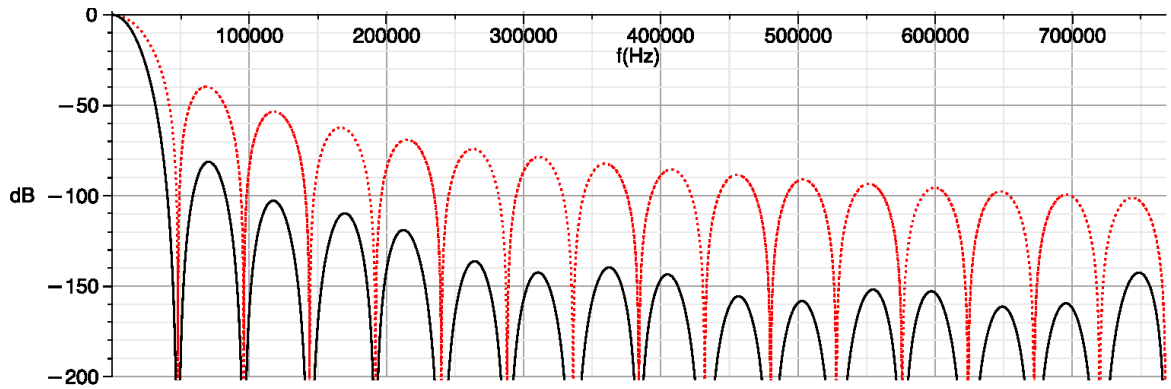

Figure 8 Ultrasonic sensitivities at 1x output rate.
Red: minimal recipe [3, 3, 3, 3, 3, 3, 3, 3, 3]; Black: 'safe' recipe [3, 3, 3, 4, 4, 5, 5, 6, 7]

The greatest vulnerability when downsampling to 2x is probably 81 kHz aliasing to 15 kHz, where the alias suppression is 45 dB.

This result is discussed later in section 9.

6.3 Quantisation-free output at 8x, 4x and 2x

The 'minimal' 1x recipe above increases the bitwidth by 27 bits, so if we start from a 24 MHz 7-bit modulator the unquantized 1x signal would occupy 34 bits and would not fit into a standard interface. For 2x we could reduce the recipe to eight '3's instead of nine and deliver 31 bits.

However, we can obtain greater flexibility by starting from 12 MHz and saving one decimation stage. We can also consider a 6-bit modulator, saving one bit throughout. If no other changes were made, the 12 MHz 6-bit modulator noise at near-audio frequencies would be 33 dB higher that the 24 MHz 7-bit example. However, by changing the Butterworth shaper to a suitable Chebyshev as shown in Figure 9, we can change this disadvantage to a 1 dB advantage and deliver a 32-bit integer result!

In Appendix 1 we tabulate some recipes and illustrate the output bit depth and noise performance obtainable by the methods shown in this paper. Appendix 2 isolates examples that can output from the decimation chain without quantisation – essentially an integer representation of the modulator output at either 24 or 32 bits.

Regarding the third question, the answer for 1x signals is probably 'no' if decimation starts from a 7-bit modulator at ~24 MHz, since the CIC chain will add 27 to the bitwidth.

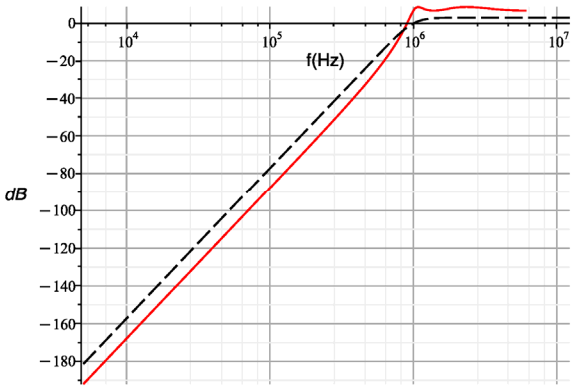

Figure 9 Black: 24 MHz Butterworth shaper; Red: 12 MHz Chebyshev shaper

6.4 Flattened Output Option

The CIC-downsampled signal inevitably has a frequency response droop, suggesting that an ADC chip should provide also an option for an output with response flattening.

The droop depends on configuration, but typically could be 0.04 dB @ 20 kHz on the 16x output. At 2x it will be nearer 2 dB @ 20 kHz. It is known however that a CIC droop can be corrected using an all-pole filter of surprisingly low order.[8] For example, a second-order filter was used to generate the 96 kHz response shown in Figure 10, which is only 0.013 dB down at 20 kHz and 0.06 dB down at 30 kHz.

The few coefficients required can be coarsely quantized so the conceptual multiplications can be expanded as shift-and-add operations and a chip need not incorporate a full multiplier.

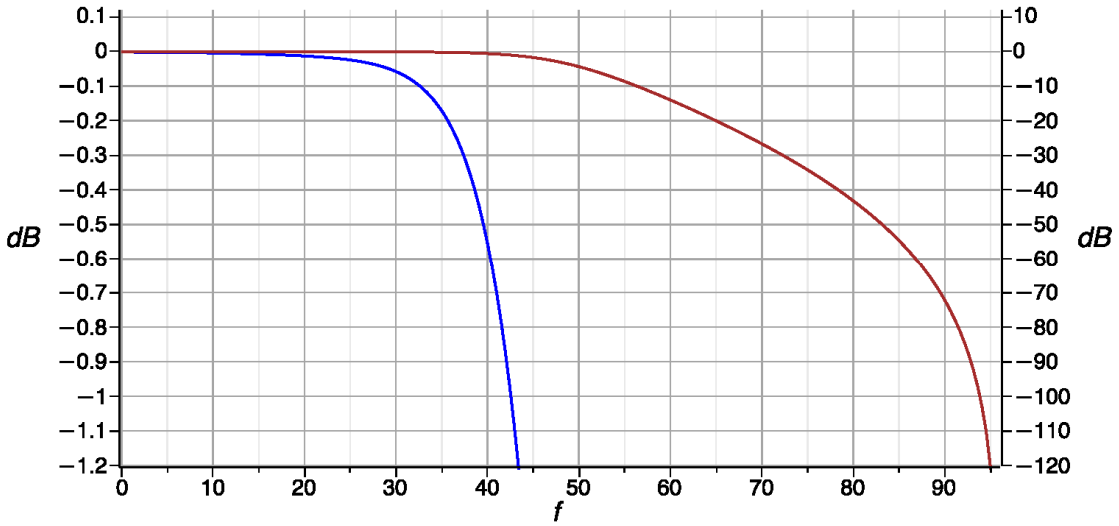

Figure 10 Flattened frequency response (kHz). Blue curve using expanded LH axis.

A signal quantization is unfortunately needed but it can efficiently include TPDF dither as the dither is not needed at frequencies higher than the output sampling frequency. Further, the quantization allows a high-quality 24-bit output at 2x, which was not contemplated in the unquantized case.

7 Performance

Figure 10 shows the flattened response of the output at 96 kHz, using a 2nd-order feedback FIR filter with 12-bit coefficients. Note that the blue trace relates to the highly expanded dB scale on the left, showing the flatness obtainable from the low-order filter.

The upper plot of Figure 11 shows the compact impulse response of this flattened output at four output rates; the lower plot compares the native and flattened responses for the 384 kHz output.

In all cases we see that the response is compact and minimum phase. Figure 12 shows the step response for the 384 kHz output, with a rise-time under 4 µs. The data for these impulse responses were measured from an FPGA implementation.

8 Optional Post Process

Should a completely ring-free response be desired, the ADC's flattened output can be postprocessed using the same coefficients as in the flattening filter, but now configured as an FIR (all-zero) filter to cancel the flattening. This possibility may be welcomed for mixing projects with the highest transparency, e.g. in a 64-bit-float workstation. [11]

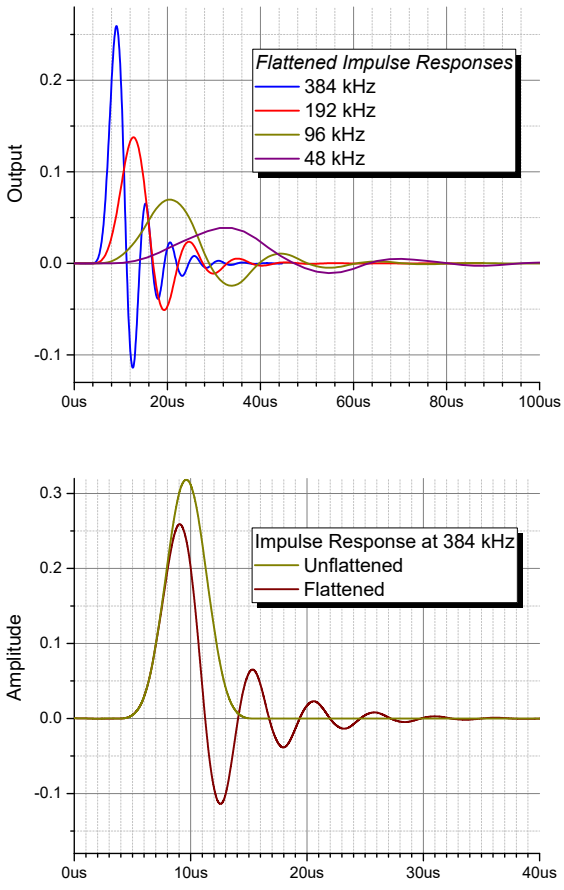

Figure 11 Showing (upper) the flattened impulse response for 1x–8x output rates and (lower) the native and flattened impulse responses for 384 kHz.

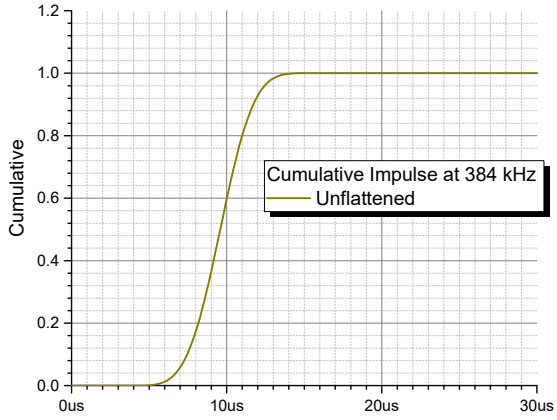

Figure 12 Showing the step response for the 384 kHz output.

9 Summary

In this paper we have illustrated that an A/D converter can be constructed using a much-simplified decimation path, based on B-splines. The time-domain performance of such a converter is exemplary, with a compact minimum-phase impulse response. This signal output can in many cases be sent over 24 or 32b interfaces without any quantisation (other than in the modulator), fully avoiding the addition of modulation-noise from filter quantization. Such an 'integer' signal should bring the highest transparency from an analogue source to a workstation that can support the precision. [11]

Depending on the output sample rate, the user may select a fully flattened frequency response (which invariably requires one additional dithered quantisation) or use the un-flattened version, e.g. for a mixing process. That the flattening can be reversed or performed downstream in a DAW, introduces scenarios where higher quality can be transported over existing interfaces.

This A/D converter with a B-spline decimation path promises high quality in all contexts and extends the hierarchical proposals in [1].

We have illustrated various possibilities for 6-bit and 7-bit modulators operating at 128x, 256x and 512x base rate.

Key differences from the more usual design are:

- The B-spline chain has lower computation complexity and can be implemented without a multiplier and on some platforms use less silicon.
- A more conventional design will use longer filters and several quantisation steps which, depending on the sophistication of the dither, add modulation noise and temporal blur.
- The design reduces noise from the modulator to levels in the audio range that will be well below analogue noise or errors from element matching in the modulator.
- A more conventional design uses sharper filters in decimation and may be more immune to ultrasonic aliasing.

To put this last point in context, it is helpful to refer to a survey of peak spectral levels in a large corpus of music in Figure 13 below, reproduced from [10].

In section 6.2, we showed that in a system using the 'gentle' recipe for a 96 kHz output, the worst-case vulnerability is for a signal around 81 kHz, creating an alias at 15 kHz reduced in level by 45 dB. Of course, there can be tonal signals in that frequency range from monitors or power supplies, but in Figure 13 we can see that the expected worst-case audio is already 80 dB below full scale, implying that in-band alias components would be below –125 dB and, significantly, 70 dB below the music content. This underlines the principles in [1] of improving time-domain performance by not using excessive filtering.

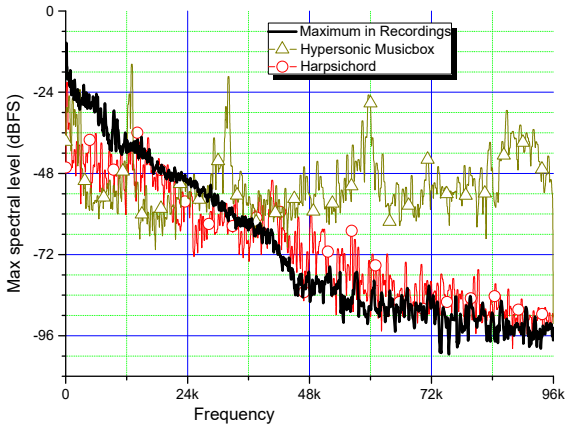

Figure 13 Showing spectrum analysis of recordings. In black is the maximum at each frequency of the peak spectra taken over a large corpus of commercial 24-bit recordings. Two unreleased stressful recordings are also included; all measured on 99% basis. Shown in red with circles is a harpsichord recording for which a wideband microphone was placed under the instrument lid, near the plectra. The olive curve with triangles shows a recording made with a similar wide-band microphone placed inside a vintage Polyphone music-box and is the most stressful recording we have discovered so far in the archive.

10 Acknowledgements

The authors would like to thank Lenbrook Industries for funding our work on this topic. We are also grateful to Malcolm Law for discussions on CIC architecture and to Alan Wood for capturing the data used in the impulse response plots.

11 References

[1] J. R. Stuart and P. G. Craven, "A Hierarchical Approach for Audio Capture, Archive and Distribution," J. Audio Eng. Soc., Vol. 67, (2019 May). Open access: https://doi.org/10.17743/jaes.2018.0062A.

[2] J.R. Stuart, and P. G. Craven, "The Gentle Art of Dithering," J. Audio Eng. Soc., vol. 67, no. 5, pp. 278-299, (2019 May.). doi: https://doi.org/10.17743/jaes.2019.0011

[3] Reiss, J.D., "Understanding Sigma-Delta Modulation: The Solved and Unsolved Issues", J. Audio Eng. Soc., Vol. 56, No. 1/2, (2008 January/February)

[4] Jackson, H.M., Capp, M.D., Stuart, J.R., 'The audibility of typical digital audio filters in a high-fidelity playback system', presented at *AES 137th Convention* (Oct. 2014) http://www.aes.org/e-lib/browse.cfm?elib=17497

[5] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling Signals with Finite Rate of Innovation," IEEE Trans.Sig. Proc., vol. 50, no. 6, pp. 1417–1428 (2007 May).

[6] Hogenauer, E., "An economical class of digital filters for decimation and interpolation", Acoustics, Speech and Signal Processing, IEEE Transactions on, vol.29, no.2, pp. 155–162, (1981 April)

[7] Unser, M. et al., "B-spline signal processing: Part I-theory", IEEE Trans. Signal Processing, vol.41, no.2, pp 821–833. (1993)

[8] L. Ferrer-Arneau, "Decimator filter based on B-splines", 19ty IMEKO TC 4 Symposium and 17th IWADC Workshop "Advances in Instrumentation and Sensors Interoperability" July 18-19 2013, Barcelona, Spain.

[9] R. Lyons, "A Beginner's Guide To Cascaded Integrator-Comb (CIC) Filters" https://www.dsprelated.com/showarticle/1337.php accessed 19th July 2024

[10] Stuart, J.R., Hollinshead, R. and Capp, M., 'Is High-frequency Intermodulation Distortion a Significant Factor in High-resolution Audio?', J. Audio Eng. Soc., Vol. 67, No. 5, pp. 310–318, (May 2019) Open access: http://dx.doi.org/10.17743/jaes.2018.0060

[11] Angus-Whiteoak, J., Stuart, J.R., "A Deep Dive into DAW Digits: Can Floating Point Save You?", *AES Fall Convention,* Streamcast, (Oct. 2021)

12 Appendix 1: Table of Configurations

All values in the last three columns relate to 6 bits from a modulator running at 12MHz.
For 7 bits, the noise reduces by 6dB and the number of output bits increases by one.

Index		Noise unweighted 0-20kHz	Noise rises above white -130dB in 20kHz	Output bits
	Configs for 16x			
A	[4, 4, 5, 5]	-206dB	129kHz	24 bits
B	[4, 4, 4, 5]	-203dB	117kHz	23 bits
	Configs for 8x			
C	[5, 5, 5, 5, 5]	-206dB	125kHz	31 bits
D	[3, 3, 4, 4, 4]	-180dB	82kHz	24 bits
E	[3, 3, 3, 4, 4]	-167dB	57kHz	23 bits
	Configs for 4x			
F	[4, 4, 4, 4, 5, 5]	-195dB	88kHz	32 bits
G	[3, 4, 4, 4, 5, 5]	-192dB	88kHz	31 bits
H	[3, 3, 3, 4, 4, 4]	-166dB	55kHz	24 bits
I	[3, 3, 3, 4, 4, 3]	-162dB	50kHz	23 bits
	Configs for 2x			
J	[3, 3, 3, 4, 4, 4, 4]	-166dB	(>48kHz)	32 bits
K	[3, 3, 3, 3, 4, 4, 4]	-160dB	42kHz	31 bits

13 Appendix 2: Table of Circumstances

Showing a summary of configurations which can deliver outputs at 24 or 32b integer from a 12 MHz modulator. The cells show the configuration from Appendix 1 and the level of unweighted output noise due to the modulator.

Output	6-bit → 24-bits	7-bit → 24-bits	6-bit → 32-bits	7-bit → 32-bits
16x	A −206 dB	B –209 dB	A –206 dB	A –212 dB
8x	D –180 dB	E –173 dB	C –206 dB	C –212 dB
4x	H –166 dB	I –168 dB	F –195 dB	G –198 dB
2x	—	—	J –166 dB	K –166 dB