Doctoral Dissertation
Doctoral Program in Electrical, Electronics and Communications Engineering
(XXXIII cycle)

Signal processing techniques to improve interpolation and modulation in audio Digital to Analog Converters

Riccardo Peloso

* * * * *

Supervisor
Prof. Maurizio Martina

Doctoral Examination Committee:
Prof. Marina Bosi, Referee, Stanford University
Prof. Sergio Saponara, Referee, University of Pisa
Prof. Paolo Crovetti, Polytechnic University of Turin
Prof. Alberto Dassatti, School of Management and Engineering Vaud
Prof. Guido Masera, Polytechnic University of Turin

Politecnico di Torino
March 11, 2021

I hereby declare that, the contents and organization of this dissertation constitute my own original work and do not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . .
Riccardo Peloso
Turin, March 11, 2021

Summary

Nowadays digital audio is ubiquitous, it can be easily acquired and reproduced with small and portable systems. The human hearing system evolved to an high accuracy in both time and frequency domains. To match this feature, a great effort is needed to design the device that converts the digital signal to its analog counterpart for physical fruition by a human listener. This system is called Digital to Analog Converter (DAC) and can be realized in many ways. Each implementation presents a slightly different sound signature and, for High Fidelity audio reproduction, this component must be able to accurately resolve the original analog signal. It can be difficult and expensive, in particular for people looking for the ultimate sound reproduction experience. Many stand-alone high-end DACs are available on the market but there is no clear winner, also due to the subjectivity of the end-user auditory tastes. It is thus important to be able to create a converter as sonically transparent as possible to let also the pickiest consumer enjoy the music content.

There are various issues associated with DAC designs. The main source of error is the analog part of the system, as the digital part exploits its intrinsic mathematical abstraction layer to reduce the error associated to its operations to the minimum. The analog part is prone to static and dynamic mismatch errors and susceptible to various noise contaminations. Fortunately, it is possible to employ clever Digital Signal Processing (DSP) algorithms to reduce the impact of mismatch and, to some extent, of noise on the analog stage. Modern Complementary Metal–Oxide–Semiconductor (CMOS) digital circuits can correct for many analog non-idealities with a relatively small resource utilization.

The most powerful idea in this field is to oversample the input signal and apply spectral shaping to error sources in order to force the error contribution to occupy a spectral zone where it can be successfully filtered out without affecting the original input signal. The quantization error, the static mismatch error, or the dynamic Inter-Symbolic Interference (ISI) error can be all shaped resorting to appropriate algorithms. This paves the way for very compact and high performance integrated circuits realizations thanks to modern deep sub-micron CMOS technological processes.

Ultimate performance has not been reached yet, for both high-end stand-alone systems and mostly-digital integrated circuits. This thesis aims to propose some novel ways to reach the state of the art for modern audio DACs.

Oversampling usually requires an high quality interpolator to increase the sample rate. The first part of the novel contributions section revolves around two new ways to perform high quality interpolation. The first one optimizes FIR filtering by adding a recursion scheme, which is particularly efficient for linear phase filter kernels that can be represented sparsely in the frequency domain. The second method works directly in the frequency domain in a block-wise operation using a novel mix of Discrete Sine Transform (DST) and Discrete Cosine Transform (DCT) to reduce border artifacts.

After that, some new modulation schemes are explored. The first one embeds the analog mismatch error shaping directly inside the main noise shaper. This extends the error correction scheme effectiveness while resorting to the classical Delta Sigma framework, which simplifies the system design. The second one works on the system stability, proposing a simple yet effective way to ensure a stable operation with a very little computational overhead. Next, a way to increase time-domain and frequency-domain signal fidelity is presented. The time-interleaving of two signals is used to compensate errors, leading to a simplified look-ahead scheme. Lastly, a mostly-digital low-power system is presented. It exploits the combination of a low-oversampling multi-level Delta Sigma modulator, Dyadic Digital Pulse Modulation (DDPM) and a shift-register-like DAC output to create a multi-level analog signal stemming from a two-level digital bitstream.

All the presented techniques are orthogonal in the solution space. They can potentially combined together to form an interesting DAC with a high quality interpolation scheme, intrinsic static mismatch error shaping, unconditional stability and high Signal to Noise Ratio (SNR).

Acknowledgements

Through the development of this work I received a lot of help and support.
At first, I would like to thank Maurizio Martina, my supervisor in this journey, for his assistance and his mentoring on digital design together with Guido Masera. I need to acknowledge also Paolo Crovetti for his invaluable expertise in the mixed-signal field. I also thanks M. Bosi, S. Saponara and A.Dassatti for their kind support to finalize and discuss this work.
I thank my colleagues, in particular German S., Maurizio C., Kristjane C., Yuri A., Matteo S., Carmine P., Luca G., Simone A., Umberto G., Gianni C., Andrea M., Andrea C., Fabio R., Stefano B., Rossana T., Luciano P., Beatrice B., Walid W., Giulia S., Giovanna T., Fabrizio R., Marco V. and all the other members of the VLSI laboratory. You were like a second family, a reason for working happily everyday.
Next, I would not have reached this milestone without the help of my beloved parents and relatives, my girlfriend Zoe, my brother Stefano and Lucia. Their precious support, in particular during the last year, made me finish this work in time despite the lockdown difficulties.
I finally want to thank my friends for all the good times I had with them, in particular Mattia P., Nicoletta S., Francesca S., Nicola G., Louis A., Luca S., Andrea DP., Andrea M., Jessica B., Luca G., Xavier D., Cecilia C., Francesco L., Alessandro C., Paolo S., Greta M., Alberto M., Alice S., Michele P., Matteo C., Ali S., Elisa L., Noah U., Daniele B., Antonello P., Roberto C., Enrico V., Giuseppe D., Anna Chiara A., Luca B. and all the others I forgot to mention. Thanks for all.

Contents

List of Figures

Acronyms

ΔΣM	Delta-Sigma Modulator
AAC	Advanced Audio Coding
ABC	Artificial Bee Colony
ADC	Analog to Digital Converter
AI	Artificial Intelligence
AKM	Asahi Kasei Microdevices
AoE	Audio over Ethernet
AoIP	Audio over Internet Protocol
AR	Asynchronous Reclock
BBDFPWM	Base-Band Distortion-Free PWM
BMC	Biphase Mark Code
CD-DA	Compact Disk Digital Audio
CIC	Cascade of Integrator-Comb
CMOS	Complementary Metal–Oxide–Semiconductor
CRC	Cyclic Redundancy Check
CSE	Common Subexpression Elimination
CU	Control Unit
DAC	Digital to Analog Converter
DAT	Digital Audio Tape
DC	Direct Current

DCT	Discrete Cosine Transform
DDPM	Dyadic Digital Pulse Modulation
DEM	Dynamic Element Matching
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DSD	Direct Stream Digital
DSP	Digital Signal Processing
DST	Discrete Sine Transform
DWA	Data Weighted Averaging
DXD	Digital eXtreme Definition
EF	Error-Feedback
FFT	Fast Fourier Transform
FIFO	First-In First-Out
FIR	Finite Impulse Response
FPGA	Field-Programmable Gate Array
GA	Genetic Algorithm
HDL	Hardware Description Language
HDMI	High-Definition Multimedia Interface
HiFi	High-Fidelity
HRA	High Resolution Audio
HS	Harmony Search
I2S	Inter-IC Sound Interface
IC	Integrated Circuit
IF	Input-Feedforward
IIR	Infinite Impulse Response
ILD	Interaural Level Difference
IMD	Intermodulation Distortion

ISI	Inter-Symbolic Interference
ITD	Interaural Time Difference
KCL	Kirchhoff Current Law
KVL	Kirchhoff Voltage Law
LSB	Least Significant Bit
LTI	Linear Time-Invariant
LUT	Look-Up Table
MAF	Moving Average Filter
MASH	Multistage Noise-shaping
MCM	Multiple Constant Elimination
MP	Matching Pursuit
MP3	MPEG-1 Audio Layer III
MQA	Master Quality Authenticated
MSB	Most Significant Bit
NLS	Non-linear Least Square
NOS	Non-OverSampling
NRZ	Non-Return-to-Zero
NTF	Noise Transfer Function
OAM	Overlap-Add Method
OF	Output-Feedback
OOB	Out-Of-Band
OSM	Overlap-Save Method
OSR	OverSampling Ratio
PCB	Printed Circuit Board
PCHIP	Piecewise Cubic Hermite Interpolating Polynomial
PCM	Pulse-Code Modulation

PGM	Pulse Group Modulation
PLL	Phase-Locked Loop
PSO	Particle Swarm Optimization
PSRR	Power Supply Rejection Ratio
PWM	Pulse-Width Modulation
RMSE	Root-Mean-Square Error
RTZ	Return-to-Zero
S/PDIF	Sony/Philips Digital Interface Format
SA	Simulated Annealing
SDR	Single Data Rate
SH	Sample-And-Hold
SMASH	Sturdy MASH
SMC	Sliding Mode Control
SNR	Signal to Noise Ratio
SoC	System-on-Chip
STF	Signal Transfer Function
TOSLINK	Toshiba Link
TPDF	Triangular Probability Density Function
TWS	True Wireless Stereo
USB	Universal Serial Bus
VLSI	Very-Large Scale Integration
VQ	Vector Quantizer
ZOH	Zero-Order Hold

Chapter 1

Introduction

Digital audio reproduction may seem a trivial procedure for those unskilled in the art. This introductive chapter aims to get a better insight into what is required to build a true state-of-the-art, high-fidelity Digital to Analog Converter (DAC) for audio applications. At a first sight, the main problems that arise when developing an audio DAC are related to the analog front-end. A plethora of issues haunt this very delicate part of the system. Through the years, researchers spent a great amount of time and resources to exploit very clever Digital Signal Processing (DSP) techniques in the digital domain to shift the analog errors contribution in the ultrasonic range of the spectrum, the Out-Of-Band (OOB) zone, in a process called error shaping.

High Fidelity audio is a journey towards perfection. Many people spend thousands of dollars in the pursuit of the ultimate reproduction system. This thesis proposes some novel ways to reach the state-of-the-art for audio reproduction, even in presence of analog sources of error.

At first a brief overview of the human hearing system, the audio signal and its digital version will be analyzed in this introductory chapter. Then, after presenting the structure of a modern Digital to Analog Converter (DAC), the features of some elements of the digital section will be discussed individually in Chapters 2 and 3. Once the underlying limitations and drawbacks of existing methods have been explained, the rest of the work will propose some novel contributions. The first part, Chapter 4, is focused on efficient and high-quality interpolation methods. The first contribution, Section 4.1, is about a novel way to implement sliding window FIR filters as the sum of truncated IIR filters

with a sinusoidal impulse response. The second contribution, Section 4.2, deals with a novel trigonometric interpolation with reduced transform-related artifacts.
The second part, Chapter 5, works on some ideas presented in Chapter 3 that have not been analyzed yet in literature. In particular, Section 5.1 shows a way to implement a multi-level signal as the sum of 1 bit contributions by better exploiting noise compensation between branches. Then, in Section 5.2, a simple way to guarantee 1 bit quantizer unconditional stability is presented by separating the quantization and the saturation non-linearities. Next, Section 5.3 exploits time-interleaved ΔΣM streams to increase output signal quality. Eventually, Section 5.4 enhance the DDPM technique with a noise shaper and an output FIR DAC.

The human hearing system

The human auditory system is a complex organ made by the physical transducers, the ears, and a powerful signal processing system, the brain [1, 2]. The ear is composed by three main parts, the outer ear, the middle ear and the inner ear. The outer ear collects the external medium pressure variation and it is tuned to approximately 4 kHz. The middle ear, composed by small bones called the hammer, the anvil and the stirrup, converts the large eardrum displacement to a small but high-force displacement for the cochlea. This mechanism also mechanically matches the impedances of the air and the cochlear fluid to minimize the energy loss. The cochlea is a spiral-shaped tunnel filled with a liquid that oscillates following the eardrum input. The vibrations excite the auditory nerves at different frequencies along the length of the cochlea. The high frequencies are detected at its beginning part, while the far neurons are excited by lower frequencies. This is often called a “frequency-space” transformation. A young human listener at its peak auditory performance is able to hear in the 20 Hz to 20 kHz range. Outside this range, the hearing system filters out the excess vibrational energy like a steep high order filter. This process is nonlinear, so the loudest sounds are less amplified than the quietest ones. Also, due to a non-uniform frequency equalization, different frequencies are perceived as equally loud even if of different physical intensity. The total resolvable audio dynamic range in humans is about 130 dB, a very wide range.

Most of the hearing magic is carried inside the brain [3]. The incoming signal is processed to reconstruct a stereo mapping resorting to complex neural networks. They exploit the difference in time of arrival between the ears, the Interaural Time Difference (ITD), and the difference of sound intensity, the Interaural Level Difference (ILD) [4]. This makes possible to detect the origin of a sound and differentiate sound sources thanks to sound transients, sudden variations in sound intensity and non-stationary waveforms. Due to the frequency-based nature of the ear, most of the transient detection is carried out by the brain. Human hearing is particularly sensible to transient timing for evolutionary purposes and this affects also music enjoyment. Using high-quality audio systems, the listener is able to detect and resolve the sound stage instruments disposition with a great accuracy.

The field that studies the relationship between the human hearing sensations and the external auditory inputs is called psychoacoustics. This science lead to many interesting discoveries, like the equal-loudness non-linear curves, the hearing threshold in quiet and the important masking phenomena that are extensively used for music compression [2].

Music is the art form that exploits sound as the medium. It is generated predominantly by musical instruments, physical or digitally synthesized, and human voice, but it is possible to integrate almost any sound source in this category. Each listener has a different subjective musical taste, which complicates the definition of what can be considered music. In this thesis the focus will be on the audio range, nominally a stereo signal that occupies the 20 Hz to 20 kHz band with about 130 dB of dynamic range.

The audio signal

Fortunately, it is possible to capture the physical sound using transducers, called microphones, to convert the physical vibrations to an electric signal. This analog signal can be processed in the analog domain or digitized with an Analog to Digital Converter (ADC) to build a more useful representation. In the digital domain, it is easy to have exact and repeatable signal processing without worrying about noise and component tolerances. At the same time, digital signal can also be stored and transmitted with ease.

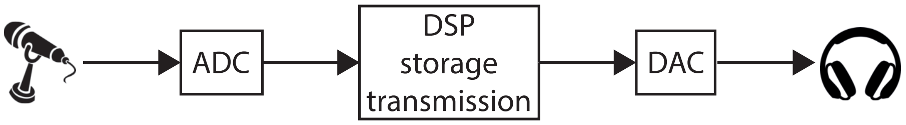


Figure 1.1: Simplified view of a typical audio recording and reproduction chain, from the source to the listener

Depending on the previous equipment accuracy, the digitized signal can carry an High-Fidelity (HiFi) replica of the sound captured by the microphone. Regarding music, many sound sources can be processed separately and mixed together by an audio engineer. Once a digital track is consolidated, it needs to be converted back to the analog domain, using a DAC, where it can be amplified and reproduced on the apposite transducer, the loudspeaker.
Figure 1.1 shows a schematic view of the recording and reproduction chain.

Digital audio

The Whittaker–Nyquist–Shannon sampling theorem [5] clearly states that the lossless reconstruction of a band-limited time-continuous signal requires a sampling frequency greater than two times the signal bandwidth. For audio, taking into account the upper hearing frequency of 20 kHz, this translates to about 40 kHz. This is the minimum sampling frequency that can accomodate the audible audio spectrum. In the real world, where brick-wall anti-imaging filters can only be approximated, it is safer to extend this frequency to at least 44.1 kHz, the Compact Disk Digital Audio (CD-DA) sampling frequency [6], or 48 kHz, the Digital Audio Tape (DAT) one [7].
These two sampling frequencies are the baseline for all the digital audio formats, with a word-width of 16 or 24 bits. Starting from this point, many digital audio formats have been proposed, with varying sampling frequencies and bit-width. They usually feature sampling frequencies that are integer power-of-two multiples of these two basic ones to reduce filter requirements and enable advanced signal processing algorithms.
As explained before, humans rely on the transient timing accuracy between the ears

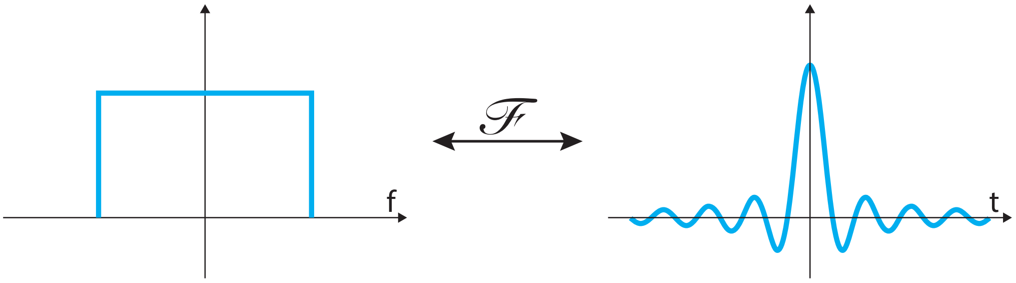


Figure 1.2: The ideal brick-wall filter requires a sinc(x) time-domain realization

to correctly reconstruct the sound stage. Using the minimum Nyquist sampling frequency, the transient timing is hidden between samples, even if mathematically all the information is still contained in the signal. The original time-continuous signal would theoretically be reconstructed by an ideal brick-wall linear-phase low-pass filter, but unfortunately it is not possible to practically realize it. This filter shows an infinite time-domain impulse response as $\mathrm{sinc}(x) = \frac{\sin(x)}{x}$ (the *Sine Cardinal* function), shown in Figure 1.2. It can only be approximated over a finite duration, for example by the Tukey-windowed low-pass filter [8], which retains the *sinc* function in the middle part of its kernel and then decays to zero to avoid filtering artifacts near the borders. Higher sampling rates would help relaxing this filter constraints, leading to a less complex system and a more life-like sound. In fact, by measuring the sound emitted by various instruments, it is possible to notice that their spectrum occupy a broader frequency range than the human maximum perceived frequency [9]. This is due to transients, as the narrower they are in the time domain, the wider will be their frequency response. The effective psycho-acousic response on human listeners is still under debat, yet a higher reconstruction accuracy in time domain can not pose any harm on the reproduction chain. It is then safe to state that already a mild oversampling (for example 96 or 192 kHz) helps reducing the requirements of the various filtering stage without adding too much complexity. These operational frequencies are still very low for modern digital circuits, and they effectively enhance the time-domain signal reconstruction accuracy.

In contrast, using even higher sampling rate, it is possible to exploit the available excess spectral space to reduce the output signal bit resolution. This will be better explained in the rest of the thesis, as this is extensively used in audio ADCs and DACs.

The digital audio signal can be stored and transmitted either uncompressed or compressed. The uncompressed version consists on raw audio signal coded as uniformly sampled Pulse-Code Modulation (PCM), with linearly uniform quantization levels [10, 11]. This digital signal can be easily processed with linear algebra through the digital part of the signal chain. It depends on two parameters, the sampling rate and the bit depth. The basic versions are the CD (16 bit at 44.1 kHz) and the DAT (16 bit at 48 kHz), but also others flavors are available and encouraged, as explained before. These signals forms the High Resolution Audio (HRA) family, with their main representative being the Digital eXtreme Definition (DXD) (24 bit at 352.8 kHz or 384 kHz [12]) and the Direct Stream Digital (DSD) (1 bit at 2.822 MHz [13, 14]). Due to some restrictions of the original DSD [15], other DSD-related formats have been proposed which either can increase the sampling rate (double-rate, quad-rate and octuple-rate DSD) and the bit depth (DSD wide [16], at 8 bit). There are still some debates about the best format, both for psycho-acoustic and implementative points of view, in particular about the required total bandwidth and the filtering complexity [17, 18].

Digital audio can be compressed, but it must be de-compressed before entering the DAC as a raw audio signal. Compression can be either lossless or lossy. Complex encoding schemes can dramatically shrink the file size at the expense of signal quality. The two most known lossy formats are the MPEG-1 Audio Layer III (MP3) and Advanced Audio Coding (AAC) [19–21]. They exploits perceptual models to decide which parts of the original raw signal can be safely discarded while retaining an acceptable output quality for a human listener. The main psychoacoustic feature employed is the frequency-domain masking, but they are also able to render transients with high fidelity using optimized adaptive time-domain overlapped window functions for filtering. Other techniques like non-uniform quantization, error shaping and the Huffman coding algorithm further enhance quality at high compression ratios.

There are many other lossy compression schemes available, but one of them seems more promising for high fidelity, the Master Quality Authenticated (MQA) technology [22]. This encoding scheme smartly compresses part of the information that usually would resides in the ultrasonic range of a HRA signal in the intrinsic noise of a 24 bit low-sampling-frequency signal (44.1 kHz or 48 kHz). Due to ADCs and hearing limitations, in digital audio only about 21 bits of the signal carry effective information, the rest of the word consists on noise. In MQA, these lower bits are filled by information that can be decoded by apposite MQA-enabled players, reconstructing the high-resolution signal. The same file can be played on a legacy player that will interpret it as a classic raw uncompressed audio signal, without resolving the additional information. There has been much debate on the effectiveness of this closed format, both for the acoustic point of view and pricing. It is important to stress out that the added high-frequency content has just a subtle effect on the hearing system but it is nevertheless useful to reduce the interpolation filter requirements.

1.1 Anatomy of an audio DAC

As the name states, a Digital to Analog Converter is composed of two main parts, the digital back-end, and the analog front-end, as shown in Figure 1.3. Analog circuitry is noise-sensitive, so it has to be electrically decoupled from the noisy digital section and usually requires a dedicated low-noise power supply. Often, externally to the DAC, a precision crystal oscillator is present. It is needed to clock both the digital and the analog parts. Usually, an internal Phase-Locked Loop (PLL) is used to derive different clock domains to accommodate the various digital stages, but the final analog stage demands a clock with high-accuracy, low-noise, and low-jitter to correctly work as desired.
This section describes the most common building blocks in a modern audio DAC system.

1.1.1 Digital Back-End

The digital back-end is the core of the system. Here the digital audio signal is processed with numeric methods to preemptively compensate for the analog stage errors. The natural way to analyze the DAC structure is to follow the input signal propagation to focus on each block.

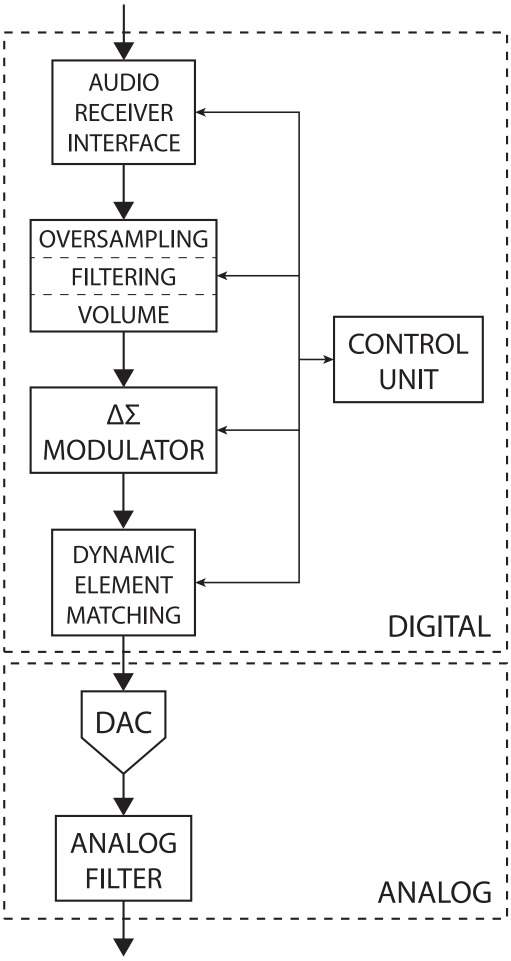


Figure 1.3: Block scheme of a digital to analog converter

Control unit

The whole digital part is coordinated by a Control Unit (CU). The CU usually also acts as an interface between the DAC and the user, receiving commands, performing actions, and returning status information. It can be built in various ways depending on the required features, like programmability, size and speed.

Audio receiver interface

The audio digital serial interface receives the uncompressed digital audio signal. The two main used consumer interfaces are the Inter-IC Sound Interface (I2S), a synchronous interface with a dedicated clock wire [23], and the Sony/Philips Digital Interface Format (S/PDIF), which embeds the clock in the signal thanks to a Biphase Mark Code (BMC) technique [24]. The latter requires a lower amount of wires, is standardized and is more difficult to handle as it sends also metadata and various information in the serial stream, codified in 32 bit wide frames. These two important interfaces are better described in Appendix B. Apart from these two protocols, it is possible to transmit uncompressed audio also over Universal Serial Bus (USB) or using Audio over Ethernet (AoE) or Audio over Internet Protocol (AoIP) [25] by suitably coding the audio stream in the payload of the selected interface. S/PDIF has also a professional version called AES/EBU, developed by the Audio Engineering Society (AES) and the European Broadcasting Union (EBU). Other multimedia interfaces can be used for uncompressed audio, like High-Definition Multimedia Interface (HDMI), but they are less used in consumer audio-only equipment.

Preprocessing

After the stream is successfully received and decoded, it is internally converted to a parallel PCM signal. The parallel stream may enter a First-In First-Out (FIFO) memory that acts as a synchronization buffer to decouple the DAC from the signal source. DSD stream can be used instead of PCM, as DSD is just a low bit-width, high-frequency PCM with an aggressive noise-shaping applied. DSD can be reproduced “as is” on a suitable reproduction system or re-converted to standard PCM for easier handling and processing. For example, it is not possible to digitally modify the amplitude of a DSD-codified signal, only an analog attenuation is possible.

Next, some first basic digital processing is done exploiting the fact that usually this is the section of the DAC that has the lowest sampling frequency and the highest bit depth. Usually, digital volume control, soft muting and basic filtering are performed at this stage. The digital volume control simply linearly modifies the amplitude level of the signal, often resorting to a soft time-based function that avoids abrupt level changes. The main drawback of this operation is that it degrades the SNR because the final noise floor remains more or less constant while the signal amplitude is here reduced. The soft mute has the same working principle, by zeroing the signal amplitude. Basic filtering can then be needed to pre-compensate issues that will arise in the next stages and to apply specific DSP functions, like pre-emphasis, equalization, harmonic distortion pre-compensation or cross-over filtering. Ideally, from the digital signal point of view, the best place to apply these functions would be just before the modulator, but the higher sampling frequency requires more computations and so usually this is performed in this part of the DAC, just before the re-sampling filter.

Interpolation

The next part of a modern DAC is the re-sammpling filter (also called the interpolation filter), that increases the sampling rate to a frequency that can be used to exploit noise-shaping and, thus, reduce the analog requirements. This process, called oversampling, is a critical part of the reproduction chain and will be better discussed in Chapter 2. Its complexity is often neglected by manufacturers because a high time-resolution interpolation requires long and complex linear-phase filters. HRA streams have less stringent requirements, as the high-quality filtering should have been carried off-line by the recording studio. Multiple designs are based on this assumption, yet it is not always ensured. Often, the minimum length filter that achieves the desired interpolation is used to reduce computations, as traditionally the frequency response is considered more important than the time response. As stated before, the *sinc* function would be the optimum filter to reconstruct the signal in the time domain but, due to its infinite-support nature, it has to be approximated for a practical implementation.
Usually, a multistage filter is employed: the required sampling frequency is reached in more than one step by using a cascade of simpler filters.

Delta-Sigma Modulator

Once the signal is interpolated to the required frequency, it can be processed by the digital $\Delta\Sigma$ modulator, which will be better discussed in Chapter 3. The modulator has the important duty of lowering the digital word width by forcing the produced quantization error to occupy mainly the OOB zone, freeing the audio baseband. This stage is the core of modern audio DACs as its output is what makes it possible to use only a small number of analog levels, reducing the analog stage requirements. With today's Complementary Metal–Oxide–Semiconductor (CMOS) transistor technology, high-frequency digital circuits with a very high integration density can be easily engineered and built. Complex and precise digital circuits can be integrated into a small silicon area. The implemented functions would be unrealizable in the analog domain, due to a lack of accuracy and noise contamination.

Most of the time ΔΣ modulators are realized as closed-loop feedback systems, more rarely as look-ahead structures. The former ones are straightforwardly implemented like digital filters. The latter ones, instead, require more computations as they have to look in the possible future horizon what will be the stream of output words that will lead to the lowest error in the base-band, while avoiding instability. The number of computations grows exponentially with respect to the look-ahead time window so it is difficult to fully exploit this technique in real-time circuits. Look-ahead ensures that the obtained signal is the best one achievable in that time window. It is usually performed off-line by software implementations, often approximated by relying on different heuristics. Some DSD tracks are encoded resorting to look-ahead modulation [26].

A common problem that afflicts ΔΣ modulators is the loop feedback stability because a non-linear saturation operator is part of the loop. Most of the high-SNR modulators are only marginally stable, which means that a control system is needed to avoid catastrophic system instabilities.

Currently, there are no analytical ways to predict a high-order delta-sigma modulator stability and so the designer has to resort to extensive simulations to ensure the correct behavior of the modulator. Even if it seems stable after simulations, it is important to still add instability detection and compensation techniques to be sure of its behavior on the field. This negatively affects the SNR, as the recirculating error is tamed by reducing the noise shaper capability to a lower performance one with increased stability. Due to the nature of audio signals, which present low average amplitudes and often a spread-spectrum behavior, the instabilities should happen rarely. The best-case scenario is a system free of instabilities, because the stability compensation techniques tend to ruin the output signal quality by changing dynamically the shape of the modulator noise. This compensation, in turns, creates unwanted intermodulation artifacts in the signal band. Additionally, compensation is difficult to ensure without limiting the maximum amplitude of the modulator input signal, either by attenuation or by signal clipping.

This is the main reason why DSD signals have a -6 dB maximum allowed amplitude under the equivalent theoretical PCM full-scale amplitude.

The digital in-band noise floor, after modulation, can reach a very low value. It is possible to design noise shapers whose limitation is the numeric precision of the digital loop filter. On the contrary, the noise floor after the analog stage, in particular due to mismatch errors and thermal effects, can get as low as -150 dB in very accurate DACs and this can be difficult to measure. These measurements require very accurate and expensive audio analyzers, usually found only in specialized laboratories. A modulator with a total in-band digital error lower than the one achievable by the physical analog stage can be said to perform well enough to be considered sonically transparent for the high-end audio sector.

As stated before, the time-domain response of the modulator is as important as the frequency-domain one in order to reproduce the transient as accurately as possible. A problem with feedback modulators like ΔΣ ones is that the error is corrected in a sample-by sample basis. Look-ahead modulators, on the contrary, predict the path that minimize the error, effectively starting to compensate its effect as soon as possible. It would thus be important to employ some look-ahead techniques for the best time-domain signal reconstruction but this is very difficult to achieve in real-time.

Another issue in the modulation part of a modern DAC is the correlation between the input signal and the quantization noise. The quantizer is often modeled as an additive source of Gaussian noise to exploit linear techniques, even if often turns to be a poor assumption. In particular, this is not a good model for a two-levels quantizer [15]. A random noise can be applied before the quantizer in a process called dithering [27] to partially de-correlate the quantization error from the input signal. If this random source is inserted in the noise shaper loop just before the quantizer, its contribution will be shaped alongside the quantization noise. The problem with 1 bit quantizers is that the required level of dither to de-correlate the quantization error from the input signal would overload the quantizer, leading to a dangerous instability, while severely limiting the input signal dynamic extension. Only partial dithering can be applied, which means that for 1 bit modulators it is not possible to completely de-correlate the quantization noise from the signal with linear additive dither. It is thus not possible to completely remove unwanted signal-related audible artifacts (in particular harmonic distortion and spurious tones). This issue can be solved by using multi-level modulators, but unfortunately the two-levels modulator is the only one inherently linear in the analog domain implementation.

Dynamic element matching

If a multi-level modulator has been used in the previous stage, an additional stage is needed to reduce the multi-level to a sum of (possibly weighted) 1 bit contributions to make the signal transfer function as linear as possible. This part of the system is called *scrambler* or Dynamic Element Matching (DEM) block [28–36]. DEM is performed in the digital domain by resorting to specialized functions. The basic scrambler aims to de-correlate each two-levels noise output, so that the error contribution due to subsequent analog mismatches becomes just an additive random noise. Higher complexity DEM techniques try to shape the error associated to each element to move the noise in the OOB zone. Basic DEMs take into account only the static analog error, while advanced algorithms are able to also cope with dynamic errors, in particular ISI mismatch. These DEMs still work on maintaining the static mismatch shaping, but they also try to decorrelate from the signal the average number of transitions between the two available levels, applying noise shaping on this source of error when possible. The final goal is to have each unary (two-level) element carry the signal constructively with respect to the others and shaped errors (digital quantization plus analog static and dynamic mismatches) destructively, or at least make them as uncorrelated as possible. In other words, each element has to be treated as an idividually constrained 1 bit delta-sigma modulator. This can be done efficiently, in particular in applications when the shaping of ISI error is neglected, using algorithms like Data Weighted Averaging (DWA). For example, it can be employed efficiently when dealing only with noise, whose harmonic distortion consists on other attenuated noise. Other DEM noise shaping techniques are possible using vector quantizers [37] but, like high-order delta-sigma modulators, they tend to suffer from instability issues. Some DACs use segmentation to reduce total quantization noise: instead of only equally-weighted unary elements, it is possible to employ also heterogeneous elements if the input signal is split in the right way. Usually, the best scheme is to have the principal elements to carry only the signal and use smaller elements only to compensate for the noise generated by the first stage, but it is also possible to have the smaller elements carry part of the input signal, if needed.

1.1.2 Analog Front-End

Once the previous stage has generated the stream of unary-weighted contributions, it is time to convert it from a logic digital representation to the actual analog signal. If the implemented DEM technique performed good enough to create streams relatively insensitive to analog mismatch, it is possible to directly convert them to an electric signal. Otherwise calibration methods are needed to ensure the signal accuracy in the analog domain. Calibration is a cumbersome and delicate procedure that can be carried off-line, when the system is not used, or on-line, while the system is in use. Even with constant on-line calibration, it is difficult to match the SNR result of DEM-enhanced DACs, but a rough calibration can still help improve the output signal quality. In fact, DEM is only a palliative that, on average, removes the error from the audio signal band to move it elsewhere, but if not applied properly it can generate other errors that can modulate back into the audio band.

The conversion from the digital domain to the analog domain is usually carried out by resorting to one of three types of discrete units of information: the electric charge (for switched-capacitor DACs), the current per clock cycle (for current-steering and resistor-ladder DACs) or voltage per clock cycle (for other exotic multi-level systems, usually for direct digital power amplification or power inverters). Currents and charges are much easier to sum under the Kirchhoff Current Law (KCL) principle. Voltages, which exploit the Kirchhoff Voltage Law (KVL), are more difficult to handle. The two former methods are preferred, in particular when dealing with integrated circuits. Switched-capacitor circuits are easily integrated on silicon as they are based on the ratio of capacitance between elements, not their absolute value. They are less sensitive to dynamic errors like ISI and jitter so they are usually preferred for integration, yet they are affected by the issue that each clock cycle the capacitors have to be recharged. For switched-capacitor circuits is not possible to resort directly to techniques like Pulse-Width Modulation (PWM) to reduce the global transition rate. On the other hand, current-based analog implementations are better suited to handle PWM-like signals and higher OSR modulators, as, usually, their operation can be faster than the switched-capacitor counterpart. For discrete implementations, this is the preferred technology because high-precision current sources or high-precision discrete resistors are cheaper and easier to manufacture than high-precision reactive components like capacitors and inductors.

The analog front-end is severely affected by the noise coming from the digital part of the circuit, so it is important to have electrical separation of the two parts. The clock jitter directly affects the quality of the signal as any jitter will be inter-modulated in the base-band due to non-linear effects. This part requires the best precision clock available, directly coming from a high-quality crystal oscillator instead than from a PLL circuit. The PLL can be safely used to clock the digital part but a synchronization mechanism (usually a chain of flip-flops) is needed before entering the analog domain. To avoid the supply-related noise it is possible to employ differential circuits. If well implemented, this error can be greatly reduced as it will act as a common-mode noise over a differential signal.

Switched-capacitors and current-steering circuits usually rely on operational amplifiers to make sure that the output signal behaves as a stable low impedance voltage source. A high-quality, low noise operational amplifier is needed to ensure the best output possible.

1.2 Oversampling and error shaping

The theoretical background behind Delta-Sigma re-quantization method will be briefly explained in this section, but a more in-depth explanation can be found in [38]. Instead of using the minimum sample rate dictated by the Nyquist-Shannon sampling theorem, it is possible to *oversample* the signal at a higher frequency, which usually is a power-of-two-times the basic frequency to ease the interpolation and the decimation processes. The memory elements inside the filters are needed to exploit the time as an additional degree of freedom, to compensate the intrinsic quantizer error in a sample-by-sample basis. A coarser quantization is necessary because analog ADCs and DACs are limited to only a few bits of linearity due to element matching. Just a 0.1% of mismatch error drops the system linearity to about 10 bits of effective resolution. This technique can also be used in the digital domain to reduce the bit-width of operands, shrinking the

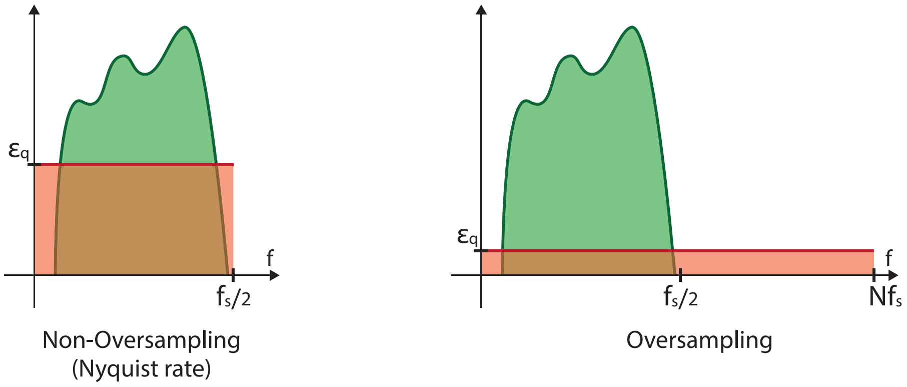


Figure 1.4: Oversampling reduces the mean quantization error ϵ_q

total required logic at the expense of a higher operational frequency.

The oversampling is needed to create a free spectral space where the excess quantization error can be stuffed. Even without noise-shaping, this error tends to occupy the whole available spectrum, as shown in Figure 1.4, reducing the average value compared to the non-oversampling case. This translates to a higher signal-to-noise ratio even without a noise shaper, but a much higher SNR can be achieved if one is used.

By defining the OverSampling Ratio (OSR) as the ratio between the actual sampling frequency and the non-oversampling *Nyquist* frequency (two times the signal bandwidth B),

$$\mathrm{OSR} = \frac{f_s}{2f_B}, \tag{1.1}$$

the quantization noise power, compared to the one in the Non-OverSampling (NOS) case, will be reduced on average by a factor equal to the OSR. The SNR will thus increase as the square root of OSR. For example, an OSR of 256 will lead to a SNR gain of 16 dB, about 4 bits of resolution, over the whole spectrum. With noise shaping it is possible to increase the in-band SNR even further at the expense of an OOB noise increase.

Another effect of oversampling is that the images of the signal, periodic repetitions in the frequency domain created by the sampling process, will be well separated. The requirements of the filters dedicated to anti-alias and anti-image can then be relaxed. Additionally, oversampling ADCs show the interesting property of capturing also parts of the signal that are out of the nominal baseband, in the ultrasonic range. It can be useful to correctly reproduce transients without time-domain smearing.

If not needed, this upper part of the spectrum can be effectively filtered out by digital filters, as they can be built with a much higher accuracy and selectivity than the analog ones. Using NOS ADCs, the unfiltered OOB signal can show as an alias in the baseband, degrading the resulting performances. To avoid this, steep analog filters would be required and they usually show a poor phase response, degrading the signal timing. Symmetrically, in NOS DACs the periodic images have to be filtered out to remove high-frequency content that could harm the rest of the reproduction chain and ruin the accuracy of transients. Also this would require precision analog filters, with the same drawbacks of anti-aliasing filters.

Overall, oversampling is a very useful tool when the final frequency is not prohibitive to be dealt with and the excess error can be efficiently removed. This is the audio case, where even an OSR of 1024 means a working frequency of about 50 MHz, which is not too difficult to be processed by modern CMOS digital designs. At such high OSR it is possible to extend the noise-free band and use it as a *guard band* to increase the cut-off frequency of the reconstruction filter, improve transient fidelity and reduce the problems related to the implementation of the modulator and the filters.

Chapter 2

Re-sampling filters

Re-sampling filters are often the most resource-consuming part of the DAC structure. Due to oversampling, the input sampling rate must be increased and adapted to the master clock frequency used by the analog section. Integer factor sampling rate converters are easier to build and this is why many high-end DACs are equipped with two high-quality crystal oscillators. One is to cope with the CD sampling frequency, namely 44.1 kHz, and the other with the DAT sampling frequency, 48 kHz. This is a sub-optimal solution as two different clock domains are difficult to handle and route, and the master clock can be one of the most expensive parts of the system, as it usually is the highest precision component. DACs often rely on the fact that this component is very accurate to reduce the requirements of the other analog circuit components. It would be thus better to accurately re-sample the source to adapt it to a unique clock frequency, not vice-versa. A non-integer re-sampling factor is otherwise difficult to handle directly (for example with the structure proposed by Farrow in [39]) but can be well approximated when oversampling the input signal.

Now that is clear that oversampling is the answer to many analog and some digital problems, it is useful to review the upsampling and interpolation scheme used by almost all oversampling DACs. The first step is to upsample the signal by inserting an integer number of zeros between adjacent samples to increase the sampling rate. This process, called *zero-stuffing*, can be performed explicitly or implicitly, like in the polyphase filter case that will be later described. This operation does not modify the baseband signal spectrum but creates periodic replicas of it to fill the newly available spectral space.

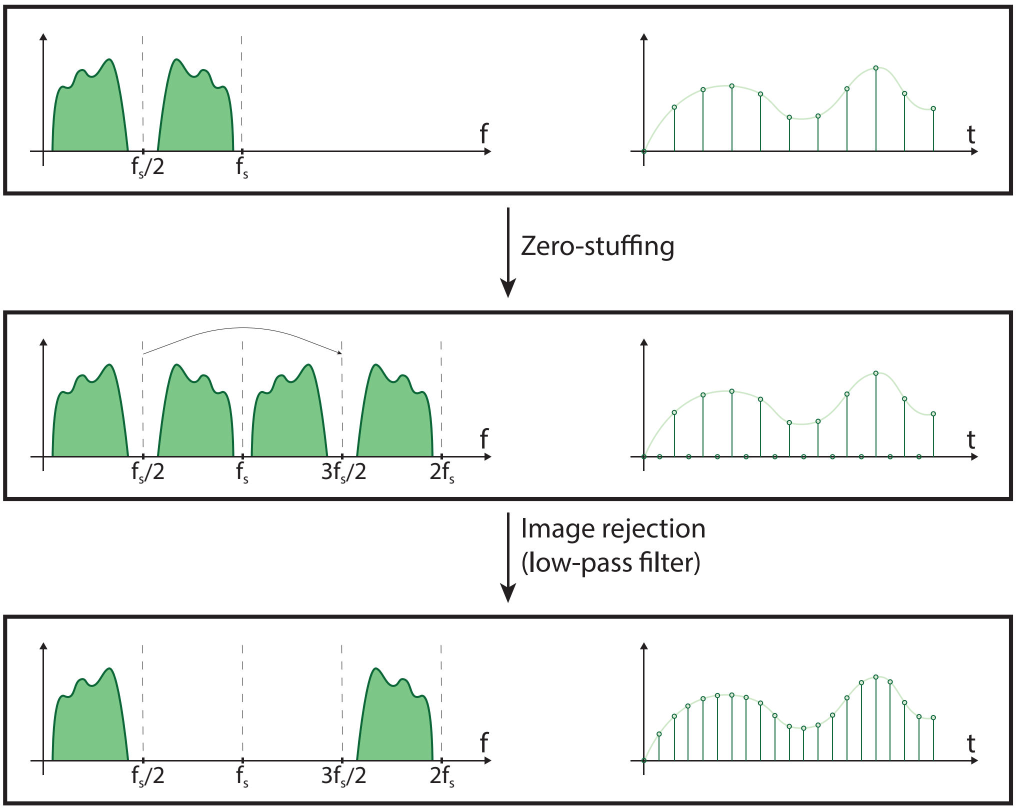


Figure 2.1: Upsampling by zero-stuffing and filtering

The second step in integer factor re-sampling is to get rid of the spectral images created by the previous step. For an audio signal, that occupies a baseband near to DC[1], it is sufficient to employ a low-pass filter called anti-image filter. Figure 2.1 illustrates this process for a power-of-two oversampling. This step is not mandatory but the high-frequency components created by the upsampling process can be problematic. They can harm the stability of the rest of the system and fold back into the audio band due to unwanted non-linear processes. It is thus important to filter them out. This filter is often called *reconstruction filter* as it recovers the original signal shape in the time domain, in particular the transient-related information used by the human hearing to accurately reconstruct the recorded track. This filter can be implemented with a filter with only zeroes in its transfer function, called Finite Impulse Response (FIR) filter, or with an

[1] The 0 Hz frequency bin is usually called Direct Current (DC)

Infinite Impulse Response (IIR) filter, which adds recursion to exploit the versatility of the transfer function poles [40]. The FIR filters are usually longer than IIR filters but they can realize exact linear-phase interpolation, which is important for proper time-domain signal reconstruction. FIR filters are commonly built without recursive structures so there are no recirculating quantization and overflow errors. If an error is made by an FIR filter, it will extinguish its contribution in a finite time thanks to its finite inpulse response property. On the contrary, IIR filters usually show lower latency due to a lower group delay, but a worse signal reconstruction fidelity due to phase non-linearity. IIR filters should be used only when the higher computational cost or the higher latency of a linear-phase FIR can not be tolerated. In this case, it is still adviseable to use filters with an enhanced phase linearity.

To reduce the implementation complexity of linear-phase filters, many efficient structures have been proposed, for example:

Polyphase filters [41–46]

This is a powerful and often exploited technique. Instead of using a single long interpolation filter working at the final sampling frequency, it is possible to avoid the upsampling-related zero-padding and directly work with a cascade of slower and shorter filters. Each stage usually performs an integer factor re-sampling, so it is possible to factorize the required OSR value to obtain the needed chain of filters.

Half-band and Nyquist filters [46–51]

These filters are very useful for integer factor interpolation as many coefficients are zeros, thus requiring no calculations when performing the convolution. The main problem with this family of filters is that the fixed -3 dB cutoff point is placed at the Nyquist frequency. They thus cannot completely eliminate aliasing or images, they have to be employed carefully. For example, if a polyphase structure is used, the first interpolation filter should not be a half-band filter to correctly remove the nearest upsampling image. Half-band filters work on half the spectrum, while Nyquist filters (also called M-th band filters) with higher integer factors.

DFT-Based Convolution [52–57]

These filters work in the frequency domain, thanks to the Discrete Fourier Transform (DFT). They return more than one sample per computation with a lighter computational workload compared to explicit FIR filtering, thanks to the Fast Fourier Transform (FFT) algorithm. The main problem is that in the frequency domain it is possible to perform only *cyclic* convolution but an FIR filter requires *linear* convolution. The border results have to be discarded accordingly to avoid cyclic convolution-related artifacts. This method is often called Overlap-Save Method (OSM) or Overlap-Add Method (OAM).

Interpolation can be also carried out efficiently in the frequency domain by zero-padding the high-frequency zone with additional zeroed bins and then performing the inverse Fourier transform. This will add a null high-frequency contribution, but the number of samples per block will be increased while leaving the original signal spectrum intact.

Interpolated FIR [58–63]

The IFIR method aims to make narrowband FIR filters as the cascade of an up-sampled FIR (which has many coefficients equal to zero) with one or more image-rejecting filters. The complexity is lower than a direct implementation but still high. The resulting filter is usually not optimized for frequency response, limiting the quality of the resulting equivalent filter.

Sharpened CIC filters [64–69]

Cascade of Integrator-Comb (CIC) filters are a particular case of FIR filters that resort to pole-zero cancellation to null the effect of a recursive structure. The basic CIC filter performs the *moving average* operation, which is a simple linear-phase low-pass filter with poor frequency response. By cascading differently weighted CIC filters it is possible to enhance the frequency response of the total structure, while maintaining the simple basic CIC structure. This type of filter is called *Sharpened CIC* filter.

The CIC often needs a fixed-point implementation to ensure the correct pole-zero cancellation and it is robust against overflow errors thanks to the modular wrap-around property of two's complement algebra implementation. CIC is often used as the last stage of a polyphase filter, where the images are already well separated frequency-wise, and its amplitude error (compared to the ideal low-pass filter) can be compensated in previous stages.

Reduced-coefficient-width FIR [70–75]

This type of filter exploits noise-shaping to re-quantize the filter coefficients and/or the input signal to reduce the word length of operands. The OOB noise generated by this method has to be compensated, for example using a CIC filter. The noise-shaping has the same issues of $\Delta\Sigma$ modulators, so it has to be approached carefully due to non-linear effects. Theoretically, it is possible to realize also *unary* coefficient filters, which have only 1 bit coefficients, and build very fast multiplication-free filters.

Shared coefficients FIR [76–80]

There are some cases in which filters present some shared coefficients. In particular, linear-phase filters have the property of *symmetry* or *anti-symmetry* in the coefficients around half the filter length. If the filter order is odd with even symmetry it is called *Type I*. If the filter order is even with even symmetry it is a *Type II* filter. If the filter order is odd with odd symmetry it is a *Type III* filter. If the filter order is even with odd symmetry it is a *Type IV* filter. Low-pass filters used in interpolation are of type I or II. By exploiting symmetry it is possible to trade half the filter multiplications with an equal number of additions. This is useful to decrease the filter area and its power consumption.

Other multiplication-free techniques [81–85]

It is possible to factorize and quantize the multiplicative coefficients to trade multiplications for sums-of-shifts (which do not require any active logic but it is only a matter of wiring adders). Common Subexpression Elimination (CSE) and Multiple Constant Elimination (MCM) algorithms can efficiently group operands to reduce the number of adders, leading to a smaller filter implementations. This is a rather inflexible structure, it is not possible to change the filter kernel without a full re-design.

Truncated IIR [86–88]

this is a particular class of filters that implements an FIR as a sum of two IIR filters, presented in [86]. The second filter is used to null the time-domain response of the first filter after a certain time delay. This is a powerful technique because it merges the advantages of an IIR filter (small size, low memory and resource-consuming) to the FIR ones (in particular phase linearity). The main problem is that this structure is subject to recirculating errors. They can exponentially grow if the IIR poles are outside the unit circle, which is mandatory for the linear-phase filter design techniques proposed in the original paper. Two TIIR filters can be used in parallel and resetted alternatively to periodically reset the recirculating error but the exponential nature of this filter limits in practice the achievable filter length. An important advantage of TIIRs is that they require only two input samples per output update, whilst the other non-recursive FIR implementations require a tapped delay line of the same length as the filter kernel. TIIRs can be implemented with a single circular buffer acting as a FIFO buffer, hence it is possible (and convenient) to use a much simpler and slower memory to store the entire delay line.

Other non-linear interpolation techniques have been proposed in the literature, which do not follow Fourier and Nyquist-Shannon theories. This is the case of *wavelet* or *spline* interpolation techniques. They try to build a good reconstruction of the signal but can affect the phase linearity [89]. It is also possible to try to enhance the interpolation with techniques labeled *super-resolution.* They guess the signal content also in the OOB zone, violating the Nyquist-Shannon sampling theorem. Recent advancements in Artificial Intelligence (AI) led to good results but it is not possible to truthfully reconstruct the original signal with this method, it will always add some artifacts. Also, due to the excessive computation required, it is difficult to perform super-resolution in real time.

Chapter 3

Re-quantization modulators

In audio applications, it is difficult to ensure the required audio quality through a low sample rate, high-resolution Digital to Analog stage. The human hearing system is very sensitive to harmonic distortion and correlated processes. Analog mismatches between elements limit the effective resolution and linearity achievable. It is then useful to employ some oversampling and re-quantization schemes to overcome these limits. Re-quantization is performed with an error-shaping algorithm, like the Output-Feedback (OF) Delta-Sigma modulator schematized in Figure 3.1, to increase the in-band SNR. Quantization error shaping is employed when the output signal word width is smaller than the input one, for example when converting a studio-quality master track to the CD format, for Analog to Digital recording, or Digital to Analog reproduction.

Error de-correlation by dithering

Each re-quantization stage adds some noise to the signal, so it is important to minimize the number of times this process is applied to reduce signal degradation. Quantization is a non-linear operation that introduces non-linear distortion. It is then important to linearize this operation to help the noise shaper remove as much error as possible. This can be achieved by a technique called *dithering*. The dither is an uncorrelated random signal that is added before quantization to make the quantizer behave more linearly, at the expense of total output signal noise. Triangular Probability Density

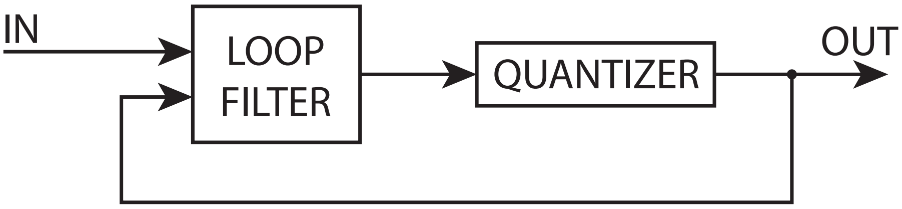
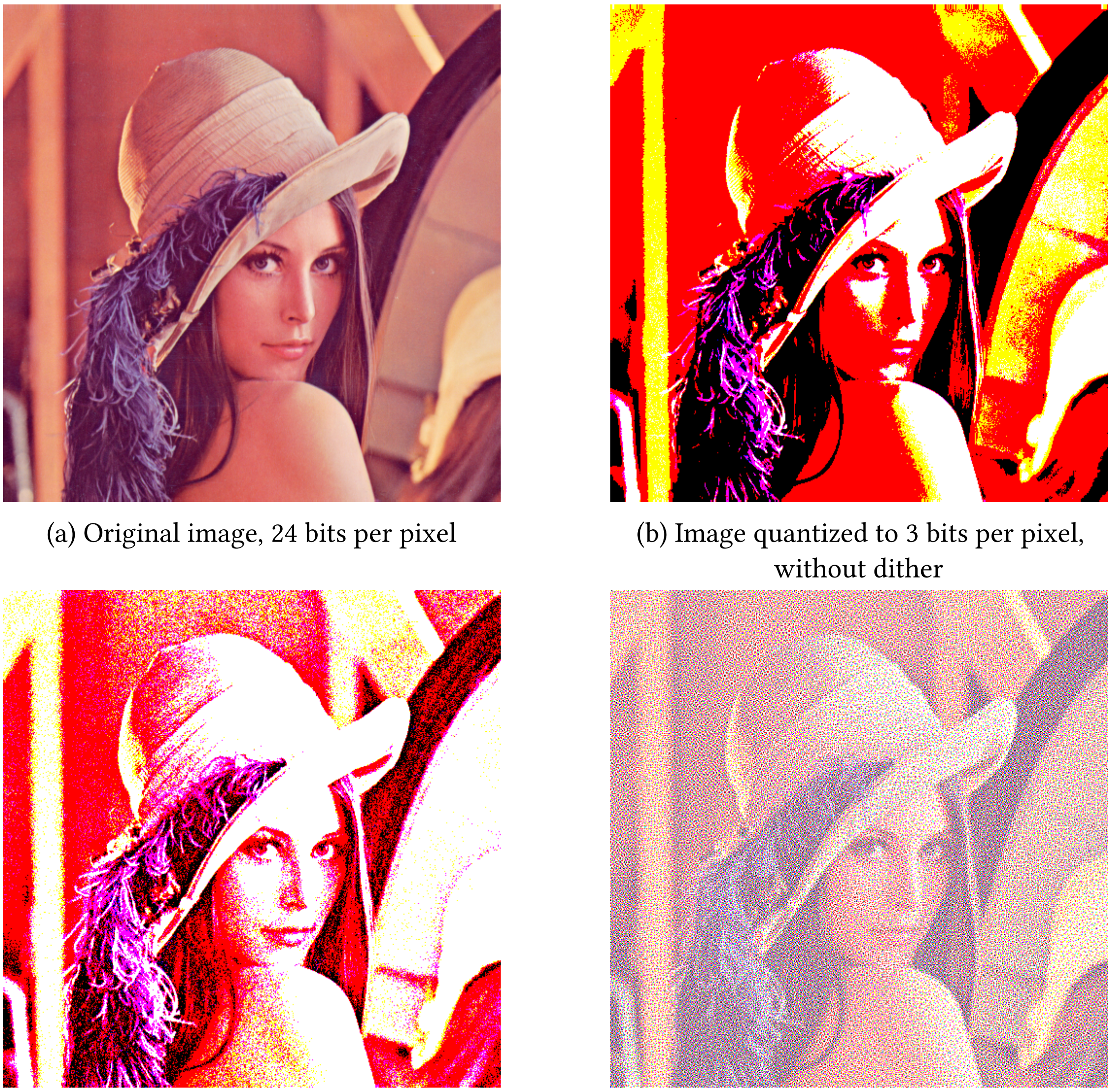


Figure 3.1: Basic output-feedback quantization error shaper (ΔΣ modulator)

Function (TPDF) noise with an amplitude of 2 times the Least Significant Bit (LSB) can de-correlate up to the second moment of the quantization error [27, 90]. Digital artifacts like harmonic distortion and spurious tones are then reduced, but the dither signal can harm the stability of the modulator and reduce the maximum allowed signal amplitude. Figure 3.2 visually shows the effect of re-quantization with and without dither for an image. The same concept can be applied to a discrete-time signal (like digital audio).

The ΔΣ Modulator

The feedback part of the noise shaper is traditionally called Delta-Sigma Modulator (ΔΣM) because, in the classical OF implementation, it takes the *difference* (Delta) between the input and the output signals and employs some *integrators* (Sigma) to analyze and correct the quantization error in a sample-by-sample basis. The signal and the shaped error are then superposed and quantized. The new output value is fed back to the Delta stage. This is called a *negative feedback* system in the Automation field. The quantization error can be de-correlated and treated like an additive gaussian source of error that has to be compensated with a closed-loop solution. This simplified linear model often offers a poor assumption of the quantizer nonlinear effects. It is possible to build both recursive and non-recursive linear filter structures but there are two main rules to observe: the first sample of the Noise Transfer Function (NTF) impulse response must be equal to one and the loop must present at least one delay. Delay-free loops are not causal and they cannot be realized in practice. In the OF topology, this means that the first sample of the noise shaper filter impulse response must be equal to zero. This severely limits the filter design possibilities. At the same time, high-order filters are more prone to quantizer overload, hence it is notably difficult to build stable high-performance ΔΣMs.



(a) Original image, 24 bits per pixel

(b) Image quantized to 3 bits per pixel, without dither

(c) Image quantized to 3 bits per pixel, with dither

(d) Image quantized to 3 bits per pixel, with dither and noise-shaping

Figure 3.2: Visual example of the effects of dither and noise-shaping on quantization

Stability

A major problem for the ΔΣM is that there is a limit to noise-shaping achievable: the re-quantization can be seen as a two-stage system, the cascade of an unbounded quantizer and a saturation element. The quantizer is just a source of noise and this does not affect the stability of the system. The saturation is what makes the problems arise. It can drive the system to instability as it prevents the feedback from correctly compensate the error accumulated by the noise shaper. Various possible solutions to this issue have been proposed in the literature, but they all require some active observation of the noise shaper status to apply non-linear compensation mechanisms. For example, it is possible to saturate or reset the values stored in the loop filter integrators to temporarily reduce the filter order to a stabler noise-shaping function, as explained in [91]. In this case the tradeoffs are a lower performance noise-shaping and complex stabilization systems for an extended input range stability.

The stability of a ΔΣM is difficult to predict exactly and extensive simulations are required to evaluate the effective design performances. If the modulator is not dithered sufficiently, the quantization noise cannot be treated as simple additive noise. The modulator will generate spurious tones and harmonic distortion. This is particularly problematic for low-resolution quantizers, for which also small dither signals can drive the quantizer to overload and the whole system to instability. The two-levels quantizer is the worst-case scenario as it is not possible to apply the required amount of dither without causing instability. The signal transfer function needs to be linearized in other ways, for example by using signal-dependent dither and instability correction schemes. Unfortunately, the two-levels quantizer is the only one guaranteed linear in the analog domain as there can be no mismatch in a single element DAC. This is why multi-level DACs are usually decomposed to a sum of many two-levels elements by DEM techniques.

When designing a ΔΣM there are various rules-of-thumb developed by heuristically observing the structure of known stable modulators or by simplifications of the ΔΣM model. Some rules dictate the sufficient conditions for a guaranteed stable modulator but they are often very conservative [92, 93]. They can mark a higher performance modulator as “possibly unstable” even if the simulations show it to be perfectly stable. The majority of modulators are stable only for a limited input signal range and only if the overload range is crossed for a limited amount of time.

One of the most common stability tests is to feed the ΔΣM with an input value with constant amplitude, starting from zero and rising gradually this value at each iteration. If that constant input does not give rise to instability in a given amount of simulation time, the modulator is considered stable for that particular input DC value and the next simulation step will be performed with a higher constant input. At a certain point, the re-circulating noise created by the bounded quantizer will overload the quantizer input, making the system unstable. Due to the nature of the integrators used by ΔΣMs, at a first sight the DC input could seem to be the worst-case scenario but it is not always the case due to the negative feedback loop effect. Particular input signals can force the ΔΣM into instability, as explained in [38]. It is not easy to predict which input pattern would make the instability rise. It is thus mandatory to simulate many possible input signals to validate the modulator behavior and always exploit stabilization techniques.

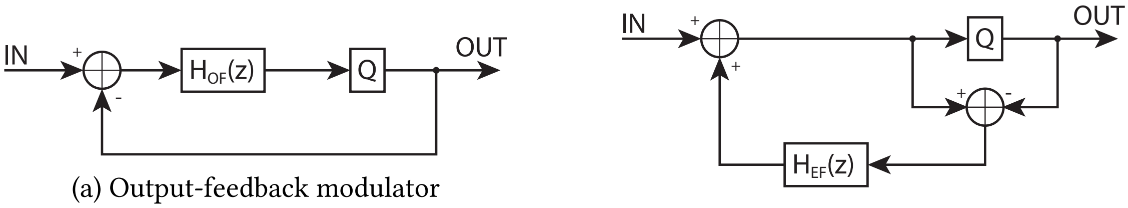
ΔΣM design is an art. It takes a lot of time and many efforts to finely tune the parameters for stability and SNR. The loop filter is usually designed starting from the desired NTF by considering the system as linear and then simulating it. If the system does not meet the required specifications it has to be re-designed with a less aggressive NTF.

Other advanced techniques employ non-linear control schemes like Sliding Mode Control (SMC) [94] or look-ahead techniques [26]. SMC is suitable for real-time modulation as it does not need a much higher computation effort compared to classic Delta-Sigma schemes. Look-ahead modulators, instead, can be effectively employed for mastering purposes but they cannot be easily used for real-time applications if more than a couple of look-ahead samples are required.

Multi-level quantizer features

Multi-level (or multi-bit) quantizers offer many advantages. Unlike the two-levels counterpart, the quantizer gain is better defined, it is easier to linearize and to ensure stability [95]. It is also possible to apply the right amount of dither without overloading the quantizer. The quantization error shows a lower correlation with the input signal due to the smaller quantizer step sizes and the presence of multiple transition regions. Multi-bit re-quantization is ideal for digital signal manipulation but there are problems with the analog stages in ADCs and DACs, as it is not possible to map a multi-level digital value to an analog level without adding static and dynamic errors. This will be better discussed in Section 3.1.1. The analog error for a multi-level converter can be shaped like an additional source of error resorting to DEM algorithms. This is a patch that follows the same rules for quantization noise-shaping. High-quality DEM circuits can be more complicated than the main $\Delta\Sigma$M itself. The amount of error that has to be moved in the OOB zone depends on the quality of the analog stage and most of the DEMs cannot take into account dynamic (ISI and jitter) errors. Only few works (for example [96–99]) propose DEM algorithms with ISI shaping. PWM-based DEM techniques are particularly robust against ISI as the average number of transitions does not depend on the input signal. Unfortunately, many ISI-shaping algorithms do not take into account that also the transition instant is crucial to get a correct shaping and still rely only on a basic transition counter. Additionally, following the analysis in [100], only one of the possible transition types is considered in these DEMs. For example single-sided PWM signals will have one edge with fixed periodic timing while the other edge will be signal-dependent. The ISI will behave differently in the two cases. Also, this approximation works only for small ISI errors and for high OSRs. A more comprehensive way to look at the ISI contribution would then consist on analyzing the spectrum of the ISI error for all the various possible output signal transitions. ISI can be reduced also in the analog section by pulse forming, for example resorting to raised cosine DACs [101] and RTZ encoding. Jitter is more difficult to deal with, as it encapsulates all the non-repetitive errors that affect signal transitions, in particular the clock uncertainties. In high-end audio equipment, this problem is reduced by using ultra-high precision, ultra-low phase noise crystal oscillators. This is expensive due to production cost, added circuit physical space, and current consumption. Like for the ISI, jitter can be shaped out by analog circuitry but, in general, it is difficult to get good results. All the non-linear errors in the system have to be compensated somewhere by high-precision components to achieve the desired linearity. Usually, the highest precision parts in ADCs and DACs are the clock source (a crystal oscillator) and the voltage reference, plus in some cases some high-accuracy resistors to modify the voltage reference value.

In the end, one of the most effective ways to reduce the jitter contribution is to have a small average amount of transitions and thus the lowest viable frequency PWM is usually a good candidate for this, but it is still difficult to reach the required performances with this encoding scheme.



(a) Output-feedback modulator

(b) Error-feedback modulator

Figure 3.3: Common Delta Sigma modulator structures

After this basic overview of the various tradeoffs in Delta-Sigma Modulators, the next sections will explain the various ΔΣM architectures more in-depth. Look-ahead and non-linear controls are out of the scope of this thesis so the reader is encouraged to look for example to [26] and [94].

3.1 Two-levels output

The 1 bit quantizer is the simplest Delta-Sigma modulator from the analog implementation point of view, but it is the most difficult to deal with at the algorithmic side. Usually, it is implemented as OF or as Error-Feedback (EF) closed-loop structures, which are both based on the negative feedback principle, as shown in Figure 3.3. In absence of additional sources of error, they are topological permutations of the same control structure, which can be generalized as state-space structures to fully exploit the automatic control framework.

While ADCs can be realized as either continuous-time or discrete-time systems, the digital nature of the DAC modulator requires a discrete-time realization. Most of the modern audio DACs are made by an oversampling digital ΔΣM re-quantizer followed by an analog stage. It converts the output of the modulator from a numeric value to a physical quantity. The analog part is usually not enclosed in the feedback loop, so it is difficult to correct non-idealities of this last stage. This step is crucial to translate the obtained high-quality digital signal to the analog domain so that the result can be used in subsequent stages as a reference high quality input. The two-levels quantizer is the basic building block of modern DACs. In fact, also in the case of multi-level modulators, DEM algorithms decompose the input multi-level signal as a sum of two-levels signals. It is thus important to fully understand the 1 bit Delta-Sigma and its flaws before proceeding to the multi-level case.

Comparison of OF and EF structures

The OF structure is the most commonly used. It can take into account also errors generated by the loop filter non-idealities, like numeric errors. It is more similar to the analog implementation of a ΔΣM for an ADC system than the EF strcture. The numeric quantization issues related to the digital implementation of the loop filter are not included in this analysis, as they are mainly related to the filter structure. It is usually built as an IIR filter organized as a cascade of integrators to reduce numerical errors.

The EF structure directly measures the error made exclusively by the quantizer stage. It can often be realized with less expensive filters but it is not usually employed in ADCs due to errors relative to analog realization imperfections. Sometimes it can be used in noise coupled ADCs [102] but it has to be enclosed in an OF structure to address the analog imperfections.

Modulator design

There are two important transfer functions in Delta-Sigma modulators: the Signal Transfer Function (STF) and the Noise Transfer Function (NTF). These quantities derive from a linear approximation of the quantizer as an additive source of white noise. This approximation is valid in particular for correctly dithered quantizers but, in practice, it is not a perfect assumption. Nevertheless, this still remains a good starting point for designing the noise shaper loop filter. Under these assumptions, it is possible to engineer the NTF as a standard discrete-time Linear Time-Invariant (LTI) filter and then apply some constraints to get the corresponding loop filter.

The first step consists in the NTF design. There are various ways to create a discrete-time filter, from the classical IIR filter design methods to more sophisticated techniques like pole-zero placement with arbitrary magnitude response and global optimization. For audio applications, the NTF is usually designed as a high-pass filter with often a zero

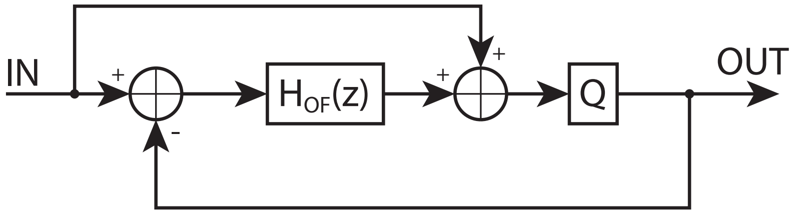


Figure 3.4: Unity STF OF Delta Sigma modulator

at DC plus some complex conjugate zeros to create distributed notches in the audio band. Poles are distributed to reduce the peak OOB noise and increase the loop stability. Usually, low-order NTFs show a faster cycle-by-cycle error correction that leads to increased loop stability while requiring less hardware to be implemented. The filter for the OF structure can be computed by re-arranging the modulator structure, leading to

$$NTF_{\mathrm{OF}}(z) = \frac{1}{1 + L(z)} \tag{3.1}$$

which can be arranged to

$$L(z) = \frac{1 - NTF_{\mathrm{OF}}(z)}{NTF_{\mathrm{OF}}(z)} \tag{3.2}$$

to find the required prototype loop filter $L(z)$. The NTF must be normalized to have unity gain at $z = \infty$ and the numerator order less or equal to the denominator order to enforce realizability [38]. The STF has no restrictions but an interesting case is when it is unity-valued for the entire spectrum, i.e. $STF_{\mathrm{OF}}(z) = 1$. This can be accomplished by the Input-Feedforward (IF) structure in Figure 3.4. IF is particularly useful when a high-quality interpolation filter has been used before the modulator, but alse because this structure minimizes the error signal that circulates in the loop filter.

The quantizer gain in the 1 bit ΔΣM has to be extracted through simulations and depends on the statistics of the input signal and often is better modeled resorting to a *quantizer transfer curve* [38]. This curve models this source of error and is linearized thanks to negative feedback. The extracted quantizer model is useful to perform a preliminary study the behavior of the loop. The root locus method [103] can provide an interesting analysis of the modulator stability.

As said before, the EF structure directly extracts the quantization error and needs a high-quality filter in the feedback path because it should not create additional error. This can be difficult to realize in the digital domain, due to finite word length effects, but is nearly impossible to realize in the analog domain to build an ADC. If the numerical error in the digital case is sufficiently low, it is possible to use simpler filters compared to the OF case. This structure is usually used for low complexity hardware implementations, in particular for the first and the second-order ΔΣM with zeros at DC. These modulators are particularly effective when used inside an OF modulator or when dealing with multi-stage modulators, which will be explained in the next section as their structure leads to a multi-level output. The EF STF is unitary by default while the NTF is simply

$$NTF_{\mathrm{EF}}(z) = 1 - L(z). \tag{3.3}$$

It is then straightforwardly possible to build the loop filter as

$$L(z) = 1 - NTF_{\mathrm{EF}}(z). \tag{3.4}$$

This noise shaper topology is usually employed to re-quantize studio master files to CD-quality through dithering and error shaping. It is easy to exploit psychoacoustic models to shape the dither noise in this configuration, thanks to the easy-to-build loop filter. In this application the filter usually consists on an FIR filter with a low quantization error. There are many possible output levels to apply dither and noise-shaping, which make a high-quality result possible.

Both OF and EF structures can be forced to have some desirable properties, like creating a PWM behavior by injecting a carrier signal in the modulator. There are some drawbacks, as these constraints usually force the system to behave differently than without the added constraint. This can seriously affect the resultant 1 bit signal quality. While the plain ΔΣM can correct an error as soon as the error is detected, a PWM ΔΣM needs to wait for the next available edge to correct for the previous error and must wait for another carrier cycle to continue the error shaping.

3.1.1 Two-levels DAC simplified model

The two-levels DAC is the basic building block of the majority of high-SNR DACs. It is then interesting to analyze this structure and build a simplified equivalent model. The main sources of error are:

- static gain error;
- static offset error;
- dynamic inter-symbol interference error;
- dynamic jitter error;
- power supply modulation error;
- other sources of error.

The reference waveform is the perfect square wave modulated by the binary bit-stream created by a two-levels ΔΣM. The binary stream can be interpreted in the analog domain by mapping the two possible values (*logical-high* and *logical-low*) to two different quantized physical values. For example voltage, current or electric charge can be used. The DAC can be unipolar or bi-polar. The difference is in offset and gain, due to a simple linear mapping difference.

Following the analysis in [104, 105], the ISI can be modeled as an additional error source that depends on the time-domain stream of bits. The first-order approximation is mainly focused on the 0-1 and the 1-0 transitions as the main source of errors but also the other types can be included in the analysis, if needed. This source of error is deterministic, as it depends on the pulse shape. It can thus be efficiently modeled. The jitter consists on random fluctuations of the reference clock period, which similarly translates to a transition timing error. Both the ISI and the jitter can be modeled as an additive source of error, whose amplitude varies on a cycle-by-cycle basis, depending on the input signal history. This model re-shape the integral error between the reference and the actual signal as an equivalent rectangular pulse area error. In this way it can be summed to the ideal waveform, as depicted in Figure 3.5. This approximation is useful for discrete-time simulations, in particular for high OSR modulators, where the transition width error

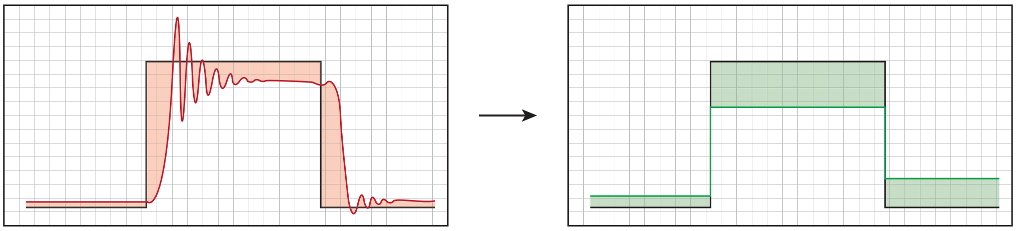


Figure 3.5: Time-domain simplification of the analog error for compact modeling purposes. The analog waveform, in red, is transformed into the green signal to maintain the bit-stream sampling frequency in the DAC model

and the clock period feature comparable durations.

Next, analog circuits are affected by power supply modulation, which can be quantified as the Power Supply Rejection Ratio (PSRR) of the system. It largely depends on the DAC typology. Usually, it can be greatly reduced resorting to differential DACs.

Other errors are due to other random or uncorrelated sources, like thermal noise or RF inter-modulation noise.

While the other error sources are difficult to model, the static and the ISI errors shows approximately time-independent behavior. Their contributions can be simulated in various ways. For example, it is possible to merge them into a single linear error contribution that depends on the last N bits. In this case, the easiest way to create a model of the DAC is to resort to a Look-Up Table (LUT) to map the 2^N possible combinations to actual output values. Usually, two bits are sufficient for modeling purposes. This compact model can be useful for calibration when the LUT is inserted in the feedback loop.

If the model has to be used to simulate the behavior of a DAC, it could be of interest to add also the other sources of error, in particular jitter. Once the static and ISI errors are known, the jitter is a quantity that approximately depends linearly on them. Jitter is mainly transition-dependent, as most of the error is produced when the signal changes polarity. If the output is composed by multiple two-levels DACs in parallel, they usually share the same master clock and so its jitter has a similar effect on all the output elements. It is important to notice that jitter sensitivity depends on the DAC typology, as explained in [106]. For example a switched-capacitor circuit is less effected by jitter in the discharge phase but the capacitors must be re-charged at each cycle, increasing the mean number of transitions. Instead, in current-steering DACs the pulse shape is more similar to the ideal digital square wave, lowering the mean transition rate.

3.2 Multi-level output

Multi-bit converters range from 1.5 bit (three levels) onward. The basic working principle is the same as the 1 bit counterpart, but with a bounded multi-level quantizer. This reduces the quantizer inherent error. It offers a better-defined quantizer gain and more levels to be occupied by the shaped quantization error. The error is better de-correlated from the signal and there are less non-linear errors. The reduced error makes it possible to use a lower OSR, which usually translates to a lower overall power consumption and a less demanding hardware. The main drawback is that it is currently not possible to directly convert the multi-level output digital signal to the analog counterpart without mismatch errors. There is the need to convert the signal to the sum of many unary (two levels) contributions, with a particular signal distribution, using a scrambler, the Dynamic Element Matching (DEM). This part of the system should be able to shape the transfer function of the analog errors, as shown in Figure 3.6. This behavior will be better addressed in Section 5.1.1. The better the unary DAC matching is, the lower will be the error that needs to be corrected by this system. The analog error can be made fairly low with well designed discrete analog stages, in particular where power and area are not issues like in stand-alone audio DAC systems. Otherwise, when the whole DAC needs to be implemented in a single Integrated Circuit (IC), which maybe must be very small and designed for low-power applications (for example for Bluetooth-based True Wireless Stereo (TWS) earphones), there could be no room for high-quality analog stages. This greatly increases the complexity of digital error shapers, both the ΔΣM and the DEM parts. In particular, for modern deep sub-micron Very-Large Scale Integration (VLSI) Complementary Metal–Oxide–Semiconductor (CMOS) technologies, the integrated transistors tends to show bad performances when working in the linear region, leading to many analog errors. Thus, for modern DAC designs, which feature much smaller transistors, the DEM has become mandatory to achieve the expected performances. At the same time, the digital part has become faster and smaller. Many digital logic gates can now be built in the area that was taken by a single high-quality

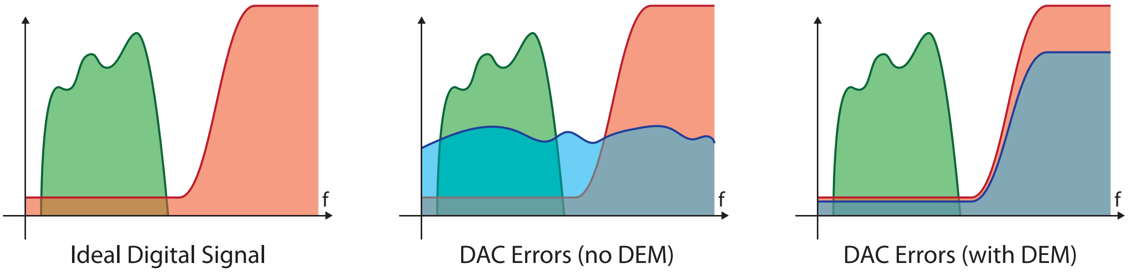


Figure 3.6: DEM principle. The analog errors (in blue) are shaped in a data-driven way

transistor in previous technologies. Additionally, modern CMOS designs can reach very high clock frequencies. It is now possible to trade analog quantization levels for time quantization levels, under the right modulation scheme.

Other big issues are cost and reliability: older technologies are now easier to access, cheaper, well optimized, and mature. Modern technologies are focused on making the transistors work as best as possible as digital switches and thus are mainly engineered for mostly-digital systems. It is still possible to use modern integration technologies like FinFET and FD-SOI techniques to build analog circuitry, but there are many problems involved [107]. An optimized solution could be to exploit modern technologies for the digital modulator and older ones for the multi-level mixed-signal part, making it a two-ICs ecosystem. Asahi Kasei Microdevices (AKM) proposed a similar solution which is composed by a digital ΔΣM (AK4191) and an analog front-end (AK4498). The two parts can be engineered and upgraded independently as long as the interface requirements are respected, paving the way for very high-quality sound reproduction and long term reliability. By using two different ICs, with dedicated power supplies, the crosstalk between the digital noise and the analog outputs is less of a concern. Unfortunately, this solution can be fully exploited only by stand-alone DAC systems due to the required silicon area.

An important fact to consider when designing these modulators is that the human hearing system is very sensitive to harmonic distortion and spurious tones. These two artifact sources have to be treated carefully, converting their energy to random noise. This is the main feature of the dithering process. The DEMs that are optimized for static mismatch errors often generates spurious tones due to ISI errors, in particular the Data Weighted Averaging (DWA) algorithm. It is necessary to shape also the ISI to avoid introducing artifacts not present in the original multi-level signal produced by the digital ΔΣM part. This is difficult to achieve, it often requires some redundancy and it reduces the maximum input signal amplitude. It is thus important to leave some quantization intervals just for noise, as quantization error shaping generates high-frequency noise that can span multiple levels, and, unfortunately, this is not achievable by 1 bit quantizers.

An alternative way to create a multi-level signal is by summing many two-levels signals. This is often adopted in multistage structures. The quantization error is refined and re-modulated one or more times to get various bit-streams that can achieve a higher-order modulation when summed together. In Multistage Noise-shaping (MASH) [108, 109] and Sturdy MASH (SMASH) [110–114] structures, the higher stability of first and second-order ΔΣMs is exploited to create many outputs that, once appropriately filtered and summed, can lead to the desired performance. An interesting property of these structures is that subsequent stages modulate only shaped error. The harmonic distortion of this "noise" is just another noise source of smaller amplitude. It has a very small tonal energy amount, which is reduced by each subsequent stage. It is also possible to apply dithering in MASH structures to additionally reduce spurious and harmonic tones [115, 116]. Multistage performances can obviously be enhanced by multi-level quantizers, there are many possible combinations but the 1 bit case remains the only one that is truly linear when converted to analog. The matching requirements between output elements is the main problem in multistage structures as they are based on subsequent error cancellation. Additionally, the first stage has the burden of carrying the whole input signal, which is the opposite of how DEM works. DEMs try to equalize the usage of elements while retaining mismatch shaping. This also means that only the first element modulates the input signal and this imbalance can reduce the effective SNR if no calibration technique is applied.

Under the presented assumptions, the ideal way to build a multi-level signal would be by directly generate static and dynamic mismatch error-aware 1 bit signals. Each output element should carry the whole signal while showing the lowest tonal error, and the error generated by each modulator should compensate for the error made by other. Each element can be independently modulated to shape its own quantization error while a global loop can take care of the sum of the elements. This would be a sort of

“DEM-in-the-loop”, highly flexible but potentially expensive in terms of computational cost. This solution would be able to ensure a high-quality bitstream for each output element, reducing in-band analog error while forcing the total output signal to behave like the output of a multi-level ΔΣM. Due to the error cross-feeding between branches, each single-element ΔΣM could potentially show higher restrictions on the total noise gain offered by the shaper. A possible solution is offered in Section 5.1.

Following this idea, it is possible to generalize this rule and split the input signal into various contributions that summed together will return the original signal. If the local error is partly compensated by the error made by the other elements, the output summation will return a multi-levels noise-shaped signal with a reduced total error. If the correlation between the initial signal and the single contributions is low, it is possible to de-correlate the analog errors from the input signal. This can lead to a lower tonal content, making the error more noise-like and thus less detectable by a human listener. This method, due to its closed-loop nature, can enclose a model of the static and dynamic error of each DAC in the loop. It would help to perform digital calibration, in particular element offset, gain, and ISI errors.

There are then many possibilities to increase performance at the modulator stage:

- faster quantization in the time domain (i.e. increasing the operating clock frequency and OSR),
- uniform multi-level quantization (each element has nominally the same weight),
- non-uniform multi-level quantization (elements can have different nominal weights, like in binary-weighted architectures. This needs accurate calibration and accuracy as it is more difficult to apply DEM techniques),
- MASH-like multi-stage architectures (the noise is re-shaped and the input signal is carried by the primary output),
- multi-stage error compensation by non-uniform weighted stages like [117] (very effective with inherent in-band mismatch shaping),
- more aggressive noise-shaping (with all the related stability problems),

- digital calibration inside the modulator loop (which requires additional circuitry),
- look-ahead modulation (very complex for effective results, good for off-line computation),
- non-linear modulator control (like SMC),
- re-quantization of multi-level signals in the time domain, notably
  - Pulse-Width Modulation (PWM),
  - Multi-level PWM (like [99] and [98], there are some restrictions on the input signal),
  - Dyadic Digital Pulse Modulation (DDPM) [118],
  - a mix of PWM and DDPM [119].

Here are listed only the most important techniques, the Delta-Sigma field is vast and an impressive number of interesting solutions have been proposed through the years. Some elements of the list are not mutually exclusive and it is possible to mix various techniques to reach the design goals.

3.3 Direct time-domain re-quantization

This section deals with open-loop solutions to trade a multi-level signal for a single two-levels stream by increasing the clock rate. Usually, the hardware required is smaller and faster if compared to closed-loop modulators, but there is no feedback compensation of non-linear errors which reduces the effectiveness of these methods. Sometimes the oversampling and noise shaping scheme is too expensive to be performed, so a direct amplitude-to-time conversion system can be a viable solution. Two interesting high-performance modulation schemes are the PWM and the DDPM.

3.3.1 Pulse Width Modulation

The PWM is a technique used to group pulses by their logic value (0 or 1). It can be generated in various ways, for example with an open-loop or a closed-loop modulation. The modulation can be *carrier-based* or *self-oscillating*. The former needs an external signal to act like a metronome, a time reference that makes the system well behaved in the time domain. The latter usually relies on a hysteresis quantizer or a dedicated non-linear control system. It avoids the use of a reference time signal and reduce the peak energy of an equivalent carrier by spreading the spectrum in the frequency domain. Both methods are well studied and understood because the reduced number of transitions of PWM signals copes well with high-power transistor switching, enabling the Class D switching mode power amplification [120–122].

The open-loop carrier-based PWM can be generated by comparing a sawtooth or a triangle signal to an upsampled and interpolated version of the input signal to accommodate the higher clock frequency. Usually, a Zero-Order Hold (ZOH) interpolation is used to simplify the structure. The open-loop PWM generation is a non-linear operation that generates a lot of harmonic and intermodulation distortion. Especially for low-frequency carriers, these errors degrade the high-quality multi-level input generated by previous stages. The non-linearity can be predicted and compensated by Volterra or Hammerstein-Wiener series-based model [122–125]. This is a difficult and resource-demanding pre-compensation method that is sometimes used to mitigate the PWM-related errors, but it destroys the simplicity inherent to the PWM scheme.

As the open-loop PWM generation is a rather interesting topic for amplification, many alternative ways to generate the output signal have been proposed under the name of Base-Band Distortion-Free PWM (BBDFPWM) [126] and Click modulation [126–131]. These modulation schemes cannot be easily performed in real-time due to complex mathematical operations and the necessity of long digital filters. In contrast, they can be viewed like look-ahead schemes because the required filters can seek in the future sample horizon for the PWM signal with the lowest in-band error. Noise-shaping is difficult to embed in these two modulation schemes, so the final two-levels stream needs a high clock frequency to make these techniques effective.

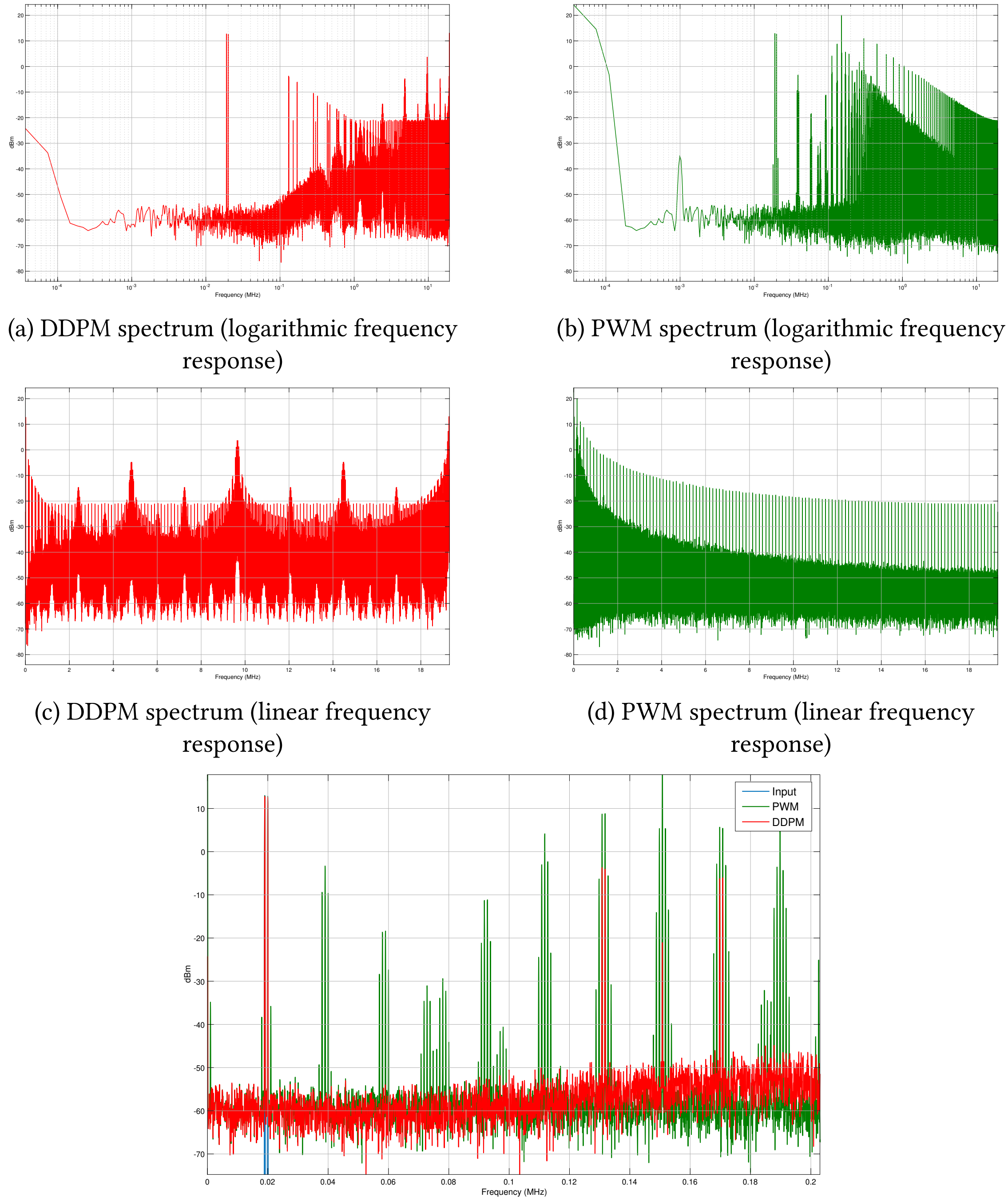
In [132] a different way to build a PWM-like signal, called Pulse Group Modulation (PGM), is presented. The output of a two-levels $\Delta\Sigma$M is grouped in temporal frames and the difference between the output of the system and the output of the ΔΣM, after a proper time delay alignment, is sent back to the ΔΣM through a secondary loop filter. This method could easily overload the main modulator due to excess loop delay and requires additional filtering to work correctly.

High SNR carrier-based open-loop PWM signal generation is then not so trivial as it could appear. The best approach is probably a closed-loop structure with a carrier frequency much higher than the maximum input signal component, with a high clock frequency and some sort of non-linearity pre-compensation techniques to make the noise shaper deal with less in-band error. A plain open-loop digital PWM generator is not suitable for high-quality audio.

3.3.2 Dyadic Digital Pulse Modulation

Dyadic Digital Pulse Modulation (DDPM) has been proposed recently by P. Crovetti in [118]. This is an open-loop re-quantization scheme that, like plain PWM, aims to reduce the output word width from a multi-level signal to a two-levels stream. It is very easy to implement as a priority multiplexer and it can work at high-speed with very low power and area requirements. The output signal, contrarily to PWM, maximizes signal switching activity to reduce the peak energy at frequencies near the base-band, which are the most afflicted by PWM. This reflects in the virtual absence of harmonic and inter-modulation distortion. In fact, at a first approximation, the base-band behavior of DDPM is equivalent to the ZOH of the input multi-level sequence. The main problems of this re-quantization scheme are tied to the analog DAC element. In particular, the output is very ISI-sensitive, and practical applications require some calibration schemes to reduce the dynamic error. The output signal is negatively affected by clock jitter and the high switching activity makes it unsuitable for high-power switching amplification. It is thus suitable only for low power analog signal generation. It is easier to filter out high-frequency noise with respect to the PWM case, as the output signal energy is concentrated far from the audio signal band. With a simple analog shift register (FIR-DAC) it is possible to build a multi-level output due to the frequency-domain alignment of the notches of the FIR-DAC with the peaks of DDPM, as will be described in Section 5.4. This alignment is valid also for PWM signals but the FIR-DAC is less effective for them, as it deals better with high-frequency peaks situations. Figure 3.7 shows the different spectra for DDPM and open-loop PWM for an input signal composed by a 19 kHz and a 20 kHz sinusoids. DDPM presents no detectable harmonic or inter-modulation distortions in the audio band, while retaining an interesting peak frequency distribution.

In [119] PWM and DDPM are used side-by-side to dither the modulated edge of a sequence of single-sided uniform PWM frames, to reduce artifacts and increase the output resolution. This is rather effective for DC generation but has not yet been analyzed for signal amplification, in particular for dynamic error mitigation. The proposed technique presents a low switching rate thanks to PWM and lowered peak noise over a plain PWM signal.



(a) DDPM spectrum (logarithmic frequency response)

(b) PWM spectrum (logarithmic frequency response)

(c) DDPM spectrum (linear frequency response)

(d) PWM spectrum (linear frequency response)

(e) DDPM shows no tonal distortion, PWM adds harmonic and inter-modulation distortions

Figure 3.7: PWM adds distortion and its carrier peaks are difficult to filter because they are too close and concentrated near the audio base-band. DDPM shows no tonal distortion and a broader peak separation

Chapter 4

Novel Contributions - Interpolation

This chapter deals with the interpolation issue. High-quality linear-phase interpolation requires a high amount of resources and so there is the need to find efficient ways to address this topic.

Two novel contributions are presented here. The first one, Section 4.1, is dedicated to long time-domain sliding window filters. The usual complexity associated with long Finite Impulse Response digital filters is greatly reduced by decomposing the filter kernel as a sum of sinusoids using a global optimization method. The result is implemented as a Truncated Infinite Impulse Response filter, lowering the number of operations and memory accesses by exploiting the FIR kernel sparsity in the frequency domain.
The second solution, in Section 4.2, directly works in the frequency domain. It exploits a combination of Discrete Cosine Transform and Discrete Sine Transform to reduce block artifacts when applying these transforms on blocks of data. These transforms and their inverses can be computed using the Fast Fourier Transform, which is well known to be efficient and is widely used for Digital Signal Processing purposes.

4.1 Sinusoidal Truncated IIR Filters

High-quality interpolation needs linear-phase filters. This is usually achieved by FIR filters, as described in Chapter 2. It is possible to exploit the time-domain cancellation property of the IIR filter response to build the Truncated IIR filters presented in [86], as schematically shown in Figure 4.1. This is, theoretically, a very powerful technique, which is also able to design linear-phase filters. Unfortunately the implementation is rather tricky as the basic behavior of the poles in IIR filters is an exponential response in the time domain, and this applies also for rounding error. IIR filter poles are placed in the transfer function by a feedback structure, and a quantizer is needed to avoid exponential internal word length growth. This means that some rounding scheme has to be applied. The rounding operation generates a rounding error, that recirculates through the IIR filter. This, in turns, reduces the effectiveness of TIIR as the tail-canceling filter could not be able to fully cancel the rounding error. The response due to the poles makes this error grow exponentially. This problem is partially solved by a periodical switch and reset between a couple of identical TIIR filters, but this still limits the maximum achievable filter length.

In this section, a class of Sinusoidal TIIR filters is presented to cope with the error growth. At first, the basic building blocks are described, and then a method to design based on Fourier Series signal decomposition and global optimization is given.

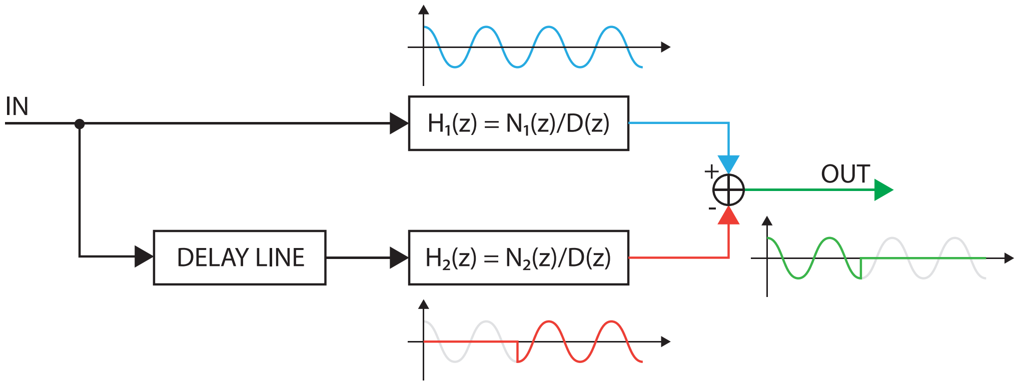


Figure 4.1: Basic TIIR diagram

4.1.1 Bounded response TIIR

To begin, two basic TIIR filters are presented here: the sinusoidal impulse response filter and the constant impulse response filter. The former presents a pair unit-magnitude complex-conjugate poles, thus it can be realized by real operations. The latter requires a single pole with unit magnitude, which is basically a discrete-time moving average filter. These two TIIRs are used as the basis to approximate the original FIR impulse response, which coincides with its kernel coefficients.

Basic sinusoidal TIIR

In [88] A. Wang and J.O. Smith, in section V.C1, introduced the *Unit Magnitude Mode TIIR filters*. Whilst the authors proposed them mainly to create polynomial windows, it is possible to extend this class of TIIR to sinusoidal impulse responses. By taking the Z transform of the generic sinusoid with frequency ω and phase ϕ,

$$\mathcal{Z}\{A_n \sin(\omega x + \phi)\} = A_n \frac{\sin(\phi) + \sin(\omega - \phi)z^{-1}}{1 - 2\cos(\omega)z^{-1} + z^{-2}}, \tag{4.1}$$

it is possible to notice that the denominator in Equation 4.1 describes a simple two complex-conjugate poles IIR resonator. Following the TIIR approach, the tail-canceling filter is another sinusoidal one of the same amplitude and frequency, but time-shifted by a z^{-L} delay. L is the length in samples of the kernel of the FIR implemented as a TIIR. The phase term of the tail-canceling filter has to be computed such that the impulse response of the total filter reduces to zero after L samples. By naming ϕ_1 the phase term of the sinusoidal IIR and ϕ_2 the phase term of the tail-canceling IIR, it is thus possible to write

$$\begin{aligned} H_{\text{TIIR}}(z) = A_n \Bigg[& \frac{\sin(\phi_1) + \sin(\omega - \phi_1)z^{-1}}{1 - 2\cos(\omega)z^{-1} + z^{-2}} + \\ & - \frac{\sin(\phi_2) + \sin(\omega - \phi_2)z^{-1}}{1 - 2\cos(\omega)z^{-1} + z^{-2}} z^{-L} \Bigg]. \end{aligned} \tag{4.2}$$

The two denominators are equal because the two IIR share the same oscillation frequency ω, reducing the required number of operations per sample.
By defining $\alpha_1 = \sin(\phi_1)$, $\alpha_2 = \sin(\omega - \phi_1)$, $\alpha_3 = -\sin(\phi_2)$ and $\alpha_4 = -\sin(\omega - \phi_2)$ it is then possible to write

$$H_{\text{TIIR}}(z) = A_n \left[\frac{\alpha_1 + \alpha_2 z^{-1} + \alpha_3 z^{-L} + \alpha_4 z^{-L+1}}{1 - 2\cos(\omega) z^{-1} + z^{-2}} \right]. \tag{4.3}$$

4.1.2 Basic constant TIIR

Like in the sinusoidal case, it is possible to derive a TIIR featuring a constant value only in the required region. This filter consists on a weighted moving average filter of the form

$$H_0(z) = A_0 \frac{1 - z^{-L}}{1 - z^{-1}} \tag{4.4}$$

which, again, shares the same $(1 - z^{-L})$ term as the sinusoidal filter in Equation 4.2. This filter is known in the literature as the Cascade of Integrator-Comb (CIC) filter [133].

4.1.3 FIR approximation

This part shows a possible way to approximate the kernel of a prototype FIR filter as a sum of sinusoids plus a constant offset. Then, a way to quantize the frequency coefficients is presented.

Impulse response approximation

A linear time-invariant system (LTI) is completely characterized by its impulse response, since it ideally contains an equal amount of all possible excitation frequencies. If two filters share the same impulse response, their behavior in the frequency domain will be the same. This is a useful property because an FIR filter can be approximated in the time domain as a sum of sinusoids and a constant offset term, as

$$y_{\text{FIR}} = A_0 + \sum_{n=1}^{N} A_n \cdot \sin\left(\omega_n x - \phi_n\right) + \epsilon, \tag{4.5}$$

where ϵ is the residual error in the approximation. This is a linear operation, leading to an LTI system.

The minimization of the L_0 norm of ϵ (i.e. the number of non-zero components) is an NP-hard problem [134, 135] so a heuristic must be used to find a sufficiently good result in an acceptable computational time.

The method proposed in this section stems from the Matching Pursuit (MP) algorithm [136]. The basic MP aims to find a sparse approximation of a signal as a weighted sum of the available basis signals collected in a dictionary, i.e. the lowest number of non-zero elements that can approximate the original function under a pre-defined error threshold. MP exploits a greedy approach to find the best basis that minimizes the error between the signal and the weighted basis. It selects at each iteration the one that best fits the residue from the previous iteration. This algorithm is proven to converge towards the exact approximation.

In Equation 4.5 there are four terms that can be optimized, namely the weights A_n, the frequency terms ω_n, the phase terms ϕ_n and the offset term A_0. The proposed greedy algorithm aims to find a single sinusoid, looking for an initial guess in the frequency domain, to fit the residual signal and then perform a global optimization for the sum of sines at each iteration, as shown in the pseudocode 1.

Algorithm 1 Sparse approximation

```
Input: FIR impulse response, error threshold, maximum terms
Output: Sparse frequency approximation
    ▷ Initialisation
residual error = FIR impulse response;
coefficients = empty;
    ▷ Iterative search
repeat
        ▷ Fourier initial guess
    take Fourier transform of residual error;
    new coefficients = offset, amplitude, frequency and phase of maximum amplitude
    sinusoid;
        ▷ Improve local solution
    new coefficients = minimize error between residual and sinusoid;
        ▷ Update global solution
    coefficients = append new coefficients to previous coefficients;
        ▷ Improve global solution
    coefficients = minimize error between FIR impulse response and global solution;
        ▷ Update residual error
    residual error = error between FIR impulse response and global solution;
until error < error threshold OR terms = maximum terms return coefficients
```

The optimization parts can be performed by a global optimization algorithm like Genetic Algorithm (GA) [137], Particle Swarm Optimization (PSO) [138], Simulated Annealing (SA) [139], Harmony Search (HS) [140], Artificial Bee Colony (ABC) [141], and many others. In this work, the global optimization is carried out by a combination of PSO for the global search and Non-linear Least Square (NLS) optimization [142] to reach the local minimum. The initial guess for the best single-sinusoid fit is based on the Fourier transform. Zero-padding the residual error can help the precision of the localization of the highest peak value in the frequency domain.

The proposed algorithm shows a sparser approximation than the basic MP, while retaining an acceptable computational time for an offline filter design.

If the filter that has to be approximated is a linear-phase FIR, it will present symmetries of even or odd type. This can be exploited in the design phase by centering the impulse response and force the optimizer to use only cosines for even symmetry (FIR type 1 and type 2) or only sines for odd symmetry (FIR type 3 and type 4). This operation lowers the number of variables, as the phase terms are no longer needed, and only half of the

kernel has to be taken into account thanks to the symmetry.
After the optimization ends, it is important to check whether the obtained result is a good approximation of the FIR filter impulse response in the frequency domain. In particular, filters with a high stopband attenuation are difficult to approximate. In this case, a direct selection of the starting frequencies for a global optimization can stem directly from the N bigger terms of the filter transfer function. This approach returns more sinusoids than the solution in Algorithm 1 but it can achieve a better approximation quality.
This technique works well only for kernels that can be approximated with a limited amount of sinusoidal terms, usually called *sparse*. For low- or no-sparsity kernels, it is not convenient to employ this method.

Coefficient quantization

Lastly, the fixed-point quantization of coefficients can be performed resorting to a similar process. For example, after the presented algorithm has reached the end, it is possible to look for the frequency coefficient that can be quantized on the available digital signal processor word length with the least amount of error. Once found, it can be fixed to that particular quantized value and then a global optimization can be run again to settle new values for the other coefficients. As before, this is a greedy algorithm helped by a global optimizer. This is particularly useful for designing the autoregressive part (the denominator) of the TIIR in Equation 4.2, which is particularly sensitive to quantization due to the recurrence needed by the time-domain realization of the filter.

Filter realization

The classic second-order IIR filter has a poor accuracy in the pole placement for low frequencies due to the $2\cos(\omega)$ term in Equation 4.2. This can be improved by different filter realizations. For example using the one shown in Figure 4.2. It still needs a single multiplication but one extra addition. In this form, the multiplicative coefficient is $4\sin^2(\frac{\omega}{2})$, which can be quantized with higher accuracy for small ω values. The roundoff noise is reduced due to the integrator blocks, as explained for similar structures in [143]. The transfer function of this structure is shown in Equation 4.6, which is the

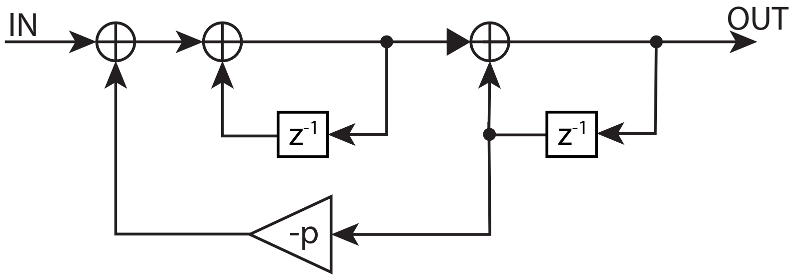


Figure 4.2: Digital resonator structure with reduced quantization error

denominator of Equation 4.2.

$$H(z) = \frac{1}{1 - (2 - p)z^{-1} + z^{-2}} \tag{4.6}$$

The complete structure of Equation 4.3, dropping the A_n weight term, can be realized as in Figure 4.3. The gray part can be shared between the parallel branches required by Equation 4.5, reducing memory requirements. The delay line can be implemented as a circular buffer because, differently from the classic FIR implementation, only the first and the last samples are needed. This greatly simplifies the structure and makes possible to easily store all the samples in an external memory.

For interpolation purposes, the part before the resonator can work at a lower frequency, namely at the input signal one, just like in a polyphase filter. This approach can lead to very high quality interpolation filters with a very limited power consumption and area utilization.

4.1.4 Example

For this example, a 1021-samples low-pass filter is approximated. It is a simple test filter created by a *sinc* function windowed with a Dolph-Chebyshev windowing function with -350 dB sidelobe attenuation (a good brick-wall filter approximation, limited mainly by the double-precision floating-point format used in MATLAB). The transfer function is designed to have a normalized cut-off frequency $f_c \approx 0.06\pi$ rad/sample.

The original kernel and its 25-terms-approximated version are shown in Figure 4.4, the

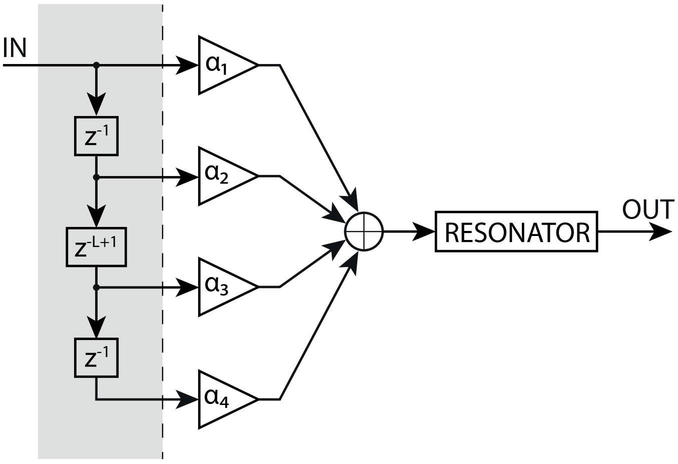


Figure 4.3: Possible sinusoidal TIIR implementation

two curves are indistinguishable at this scale, so Figure 4.5 illustrates the difference between the two. Their transfer functions are then compared in Figure 4.6.

In this case, 25 terms are not sufficient to fully approximate the input filter behavior, but the resulting filter could still be good enough, depending on the design requirements. No quantization has been performed, as this filter is designed to be used on a high-performance DSP system with double-precision floating-point support. The maximum absolute error is lower than 1.5×10^{-8} which corresponds to an FIR with about 26 bits of fixed-point coefficient accuracy. This can be also seen in the transfer function: 24 bits (typical word width of high-resolution audio) interpolation requires about 144 dB of band rejection. This filter is capable of about 150 dB and, thus, can be successfully employed in an oversampling system.

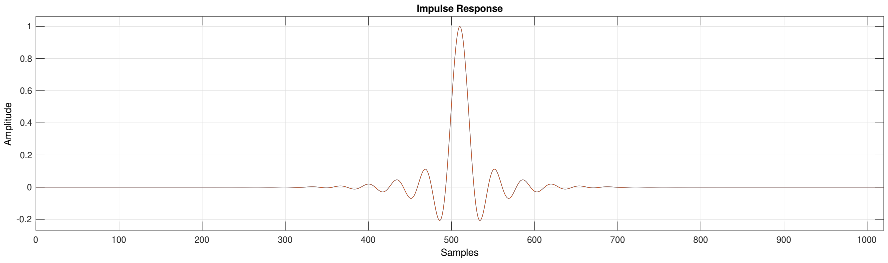


Figure 4.4: Impulse response of the original (orange) and the approximated (blue) low-pass Dolph-Chebyshev windowed filters

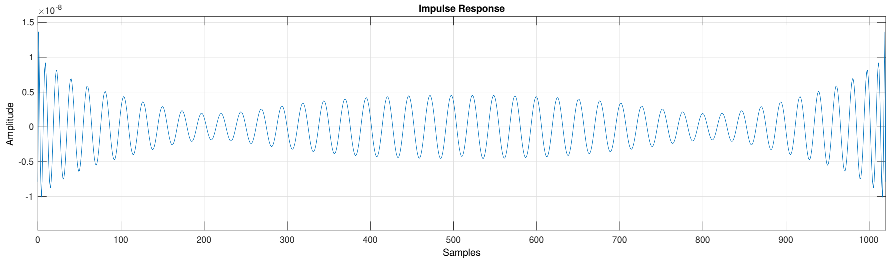


Figure 4.5: Error between the impulse response of the original and the approximated low-pass Dolph-Chebyshev windowed filters

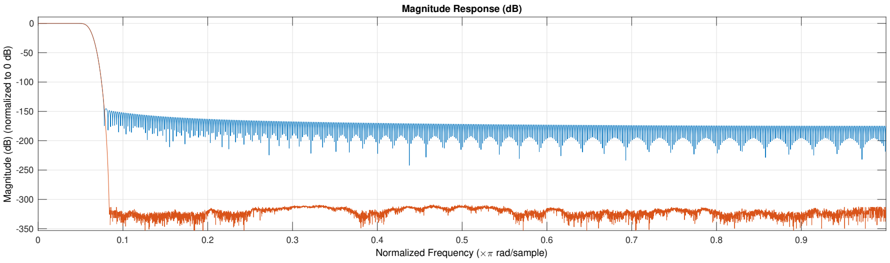


Figure 4.6: Transfer function of the original (orange) and the approximated (blue) low-pass Dolph-Chebyshev windowed filters, with normalized DC amplitude

4.1.5 Remarks

The purpose of the proposed algorithm is to approximate the kernel of an FIR filter as the sum of sinusoidal basis functions plus an offset term. A sinusoidal response filter can be constructed by an IIR filter and a finite impulse response can be obtained with the TIIR technique. The regressive nature of the IIR filter greatly reduces the computations and memory accesses required, in particular for filters that can be represented sparsely. The number of terms is optimized with a novel Matching Pursuit algorithm made by a combination of Fourier Transform, Particle Swarm Optimization, and Non-linear Least Square optimizations. Future works can focus on building a user-friendly framework to implement the proposed algorithm, also integrating coefficient quantization and direct code generation (for software or hardware designs).

The filter coefficients can be quantized with a similar algorithm by knowing the possible pole placement for a given filter realization. Due to the sinusoidal nature of the TIIR filters, it is important to employ the right structures to minimize the coefficients quantization error.

The proposed technique performs poorly on low-sparsity kernels. A classic FIR filter is advisable in this case. A typical use case scenario could be for interpolation purposes, where long linear-phase low-pass filters with a narrow bandwidth are often required.

4.2 Enhanced DST-based Interpolation

The polynomial interpolation [144–147] goal is to fit a polynomial curve through the available sampled value and then extract the required new points from the curve. Some problems arise mainly at the boundaries of the support, as each polynomial basis suppose different conditions outside the available input range. For example, the Lagrange interpolating polynomials fit a polynomial curve through the available points, neglecting the possible values of the signal outside the available support. The curve tends to diverge outside the defined interval. For real-life signals, this is usually not likely the case. Real-life signals, in this case the audio ones, often show a small mean value and bounded peak value. This is why, in particular for small signal supports, the interpolated samples can greatly differ from the ideal original values. Equidistant nodes, which is the case for discrete-time sampled signals, are proven to be difficult to handle by polynomial and trigonometric interpolation techniques [147]. They often incur in unwanted oscillations, in particular for Runge and Gibbs phenomena.

Trigonometric interpolation aims to connect the available samples using a trigonometric polynomial. Fourier series are an example of trigonometric curve fitting with orthonormal bases that minimizes the L_2 norm of the fit error. Fourier series require infinite periodic supports to work. It is possible to achieve the required support from the available finite one by infinite periodic signal repetition. The way the signal is replicated to build the infinite support defines the trigonometric basis required for fitting it. An in-dept overview is proposed in [148].

For example, the Discrete Fourier Transform (DFT) is based on the simple periodic repetition of the support. Without particular signal symmetries, the result of the DFT is a combination of sine and cosine terms (for the real formulation), or complex terms (for the complex formulation). This simple repetition leads to jumps at the support boundaries, which causes unwanted oscillations like the Gibbs phenomenon. For this reason, it is common to apply a windowing function to the input signal to attenuate these artifacts, while retaining an acceptable accuracy in the frequency domain [149, 150]. Windowing functions, however, reduce time-domain accuracy as they modify the input signal to make it behave better in the DFT framework, so they have to be used carefully in audio processing to retain the correct transient timing.

If the signal repetition is *even*, i.e. $f(x) = f(-x)$, the resulting Fourier series will present only cosine terms, leading to the Discrete Cosine Transform (DCT) [151–153]. This transform presents some interesting properties. It is real-valued, requiring only half the processing power compared to the DFT implementation, and there are no discontinuities at the boundaries due to the even symmetry. It also tends to show a greater energy-compaction in the low-frequency bins, the discretized frequency values obtained by the transform. For these reasons, the DCT is vastly employed in image and video compression as it is possible to achieve higher quantization and compression ratios [154–156].

If the signal repetition is *odd*, i.e. $f(x) = -f(-x)$, only sine terms will be present, leading to the Discrete Sine Transform (DST) [148, 151]. This transform is less used as the boundary discontinuities still are present like in the DFT case, but it still retains a good energy compaction capability. An interesting property of the DST-I is that at the boundary all the derivatives of the trigonometric interpolant are continuous, contrarily to the DCT where the function is continuous but the derivatives are discontinuous [148]. A particular case is when at the boundaries the signal is zero-valued, in this case, the odd symmetry will remove all the discontinuities. It is possible to force this condition by using DST Type I, which deliberately adds two additional zero-valued samples at the extrama of the signal to enforce this Dirichlet condition. This operation does not remove discontinuities but it just places one more sample in the middle of the jump, which helps to smoothen it and reduce the Gibbs phenomenon.

If the signal support naturally decays to zero at the borders, it would be possible to avoid the artificial zero-padding at the extrema by modifying the DST formulation. In this condition the repetition will show no discontinuities, leading to a high-quality DST-based interpolation. This section presents a simple way to enforce this condition by adaptively splitting the input signal as the sum of two contributions.

Figure 4.7 compares the three frequency transforms. Depending on the boundary conditions, there are eight types of DCT and DST. In figure only the Types II are shown for simplicity, the complete list can be found in [148].

There are many other ways to fit a curve to the available signal, like wavelet transforms, but Fourier-related transforms are usually the most employed. They can be realized efficiently and are well suited to deal with signals limited in amplitude and bandwidth,

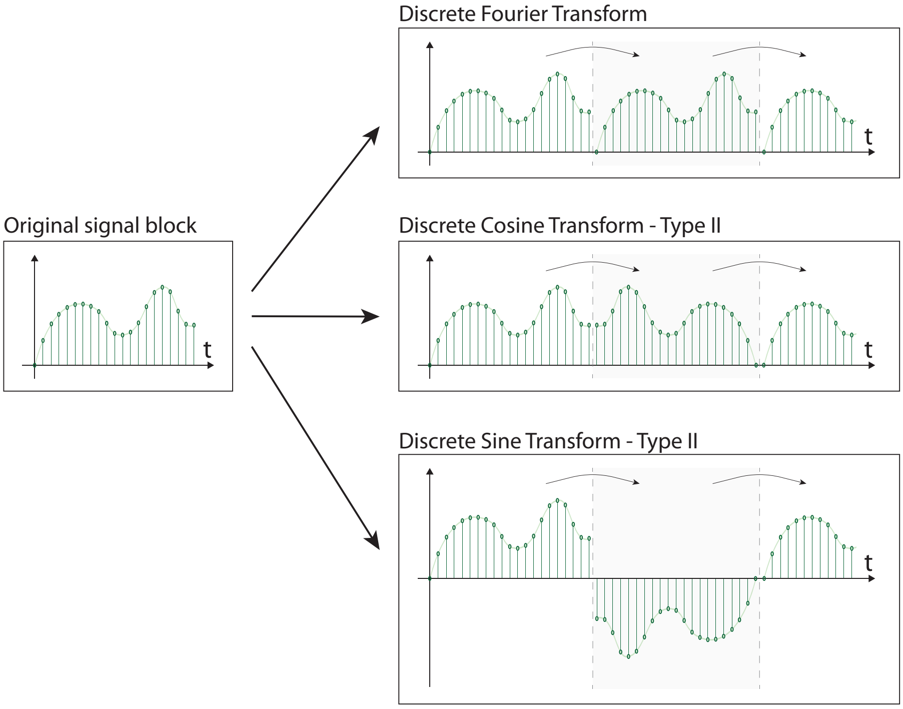

Original signal block
Discrete Fourier Transform
Discrete Cosine Transform - Type II
Discrete Sine Transform - Type II
t

Figure 4.7: Upsampling by zero-stuffing and filtering

like audio ones.

4.2.1 Derivation

As discussed before, a trigonometric curve is a good candidate to fit the available audio samples. The interpolation quality depends on the shape of the signal outside the finite length block, here called *support*, used for interpolation. Border effects negatively influence the result. Wider supports will reduce the error caused by the discontinuities in the middle of the block but they cannot completely remove this effect. Also, longer supports usually require more computational resources.

The DFT is well suited when the signal has the same value and derivative at the extrema. The DCT (mostly types I and II) when the signal has zero-derivative at the extrema, and the DST (mostly types I and II) when the extrema are zero-valued. Following this reasoning, it is possible to reduce the trigonometric fit border error by enforcing the zero-valued extrema in the DST. The fast DST algorithms would still be exploitable [157] leading to an efficient block-wise implementation.

The input signal can be decomposed as the sum of two contributions: an extrema-connecting curve and an oscillatory part. The former can be any curve that can be interpolated algebraically. For example, a cosine function with an offset term or a straight line connecting the two points. If the curve is a cosine, it could be built so that the ending points have zero-derivative, i.e. a minimum and a maximum point. This decision copes well with the statistical properties of an audio signal, which is amplitude and energy bounded like the cosine function. The latter oscillatory part is the difference, or residual, between the original signal and the extrema-connecting function.

Once the signal is decomposed in these two parts it is possible to interpolate the extrema-connecting curve with the desired method. For example if a cosine function is used, it can be interpolated trigonometrically with a single cosine term with an offset, similarly to the DCT approach. The residual part can be interpolated reformulating the DST approach taking into account that the two extrema are now exactly zero-valued and that the added samples should not be used. The cosine part will remove the repetition-induced jumps and the DST-like part will enforce the derivatives continuity at the extrema. The results of the two interpolations can be added to build the final interpolated signal, as schematized in Figure 4.8.

The first step then consists in splitting the input signal in the various contributions. Supposing the block starts at the sample $n = 0$ and ends at $n = N$, so that there are in total $N + 1$ samples, it can be written as

$$f_{\mathrm{s}}[n] = f_{\cos}[n] + f_{\mathrm{off}}[n] + f_{\mathrm{osc}}[n] \qquad \text{for } 0 \leqslant n \leqslant N \tag{4.7}$$

where $f_{\mathrm{s}}[n]$ is the discrete time signal, $f_{\cos}[n]$ is the cosine term, $f_{\mathrm{off}}[n]$ is the DC offset and $f_{\mathrm{osc}}[n]$ is the remaining oscillatory part.

Analyzing $f_{\mathrm{s}}[n]$ it can be seen that $f_{\mathrm{off}}[n]$ is simply

$$f_{\mathrm{off}}[n] = \frac{f_{\mathrm{s}}[0] + f_{\mathrm{s}}[N]}{2} \tag{4.8}$$

and $f_{\cos}[n]$ values

$$f_{\cos}[n] = \frac{f_{\mathrm{s}}[0] - f_{\mathrm{s}}[N]}{2} \cos\left(\frac{\pi}{N} n\right). \tag{4.9}$$

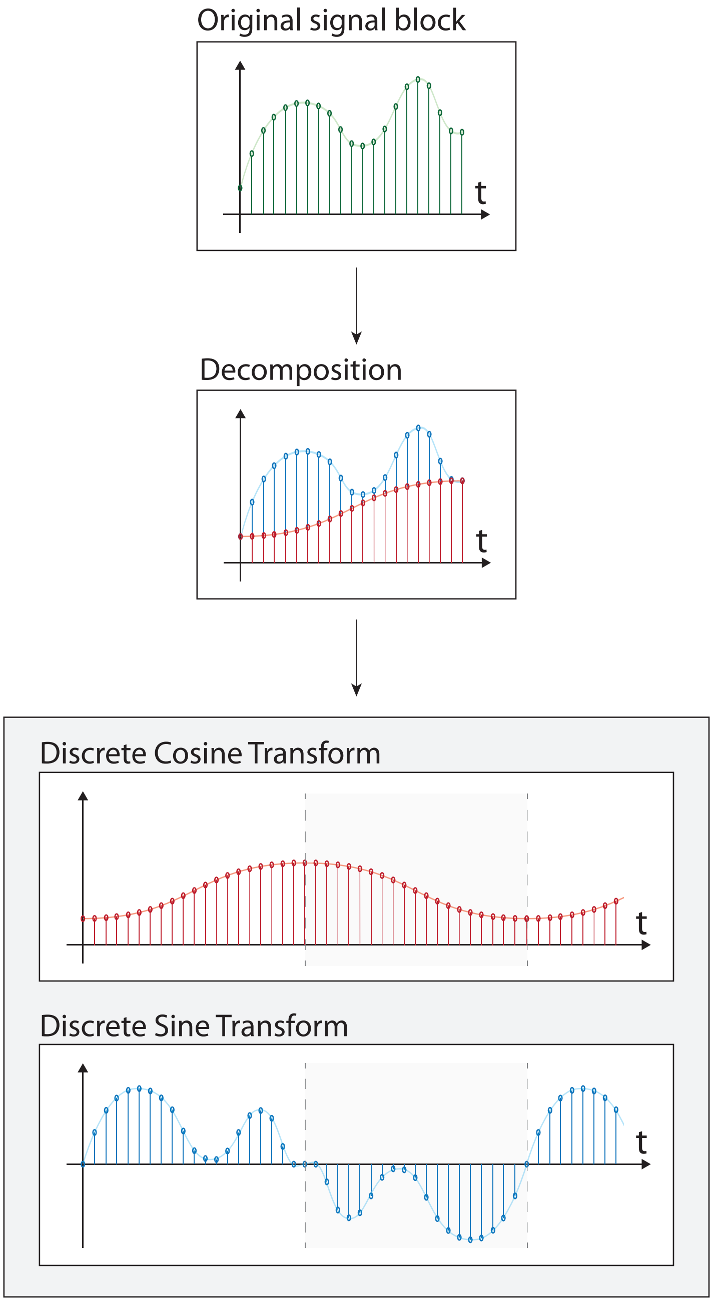


Figure 4.8: Hybrid DCT-DST interpolation

The two, combined, form an offsetted cosine that passes through the two estrema. This function is built to have zero derivative at samples $n = 0$ and $n = N$ and half cosine period equal to N samples.

These first two terms can be re-sampled straightforwardly, while $f_{\text{osc}}[n]$ needs to be interpolated following the DST type I principle. At first, $f_{\text{osc}}[n]$ is computed as

$$f_{\text{osc}}[n] = f_{\text{s}}[n] - \left[f_{\cos}[n] + f_{\text{off}}[n] \right] . \tag{4.10}$$

Now $f_{\text{osc}}[0]$ and $f_{\text{osc}}[N]$ are equal to zero and so the DST-I has to be computed only on the range $1 \leqslant n \leqslant N-1$.

The DST-I is then defined, in this scenario, as

$$F_{\text{osc}}[k] = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} f_{\text{osc}}[n] \sin\left(\frac{\pi n}{N} k\right) \tag{4.11}$$

Once the values of frequency bins have been retrieved, it is possible to zero-pad the high-frequency part of the spectrum to perform the interpolation in the frequency domain.

The new block of data will then have M bins, of which $M-N$ are zero-valued. This leads to the new samples $\hat{f}_{\text{osc}}[m]$ defined as

$$\hat{f}_{\text{osc}}[m] = \sqrt{\frac{2}{M}} \sum_{k=0}^{M-1} F_{\text{osc}}[k] \sin\left(\frac{\pi k}{M} m\right) \tag{4.12}$$

which is the same formulation as Equation 4.11 because the inverse transform of the DST-I is the DST-I itself, if the normalization factor $\sqrt{\frac{2}{N}}$ is used [153]. A further normalization by $\sqrt{\frac{M}{N}}$ is required to take into account the added points.

Equations 4.8, 4.9 and 4.12 can finally be combined to find the final interpolated values.

4.2.2 Example

In this example, a random signal source has its spectrum limited by an almost-ideal anti-aliasing filter. The resulting signal can be safely down-sampled 8 times and then up-sampled by zero stuffing to obtain a *reference* and a *down-sampled* versions of the same band-limited signal. Figure 4.9 depicts schematically the employed system, while Figure 4.10 shows the obtained spectrum and waveform.

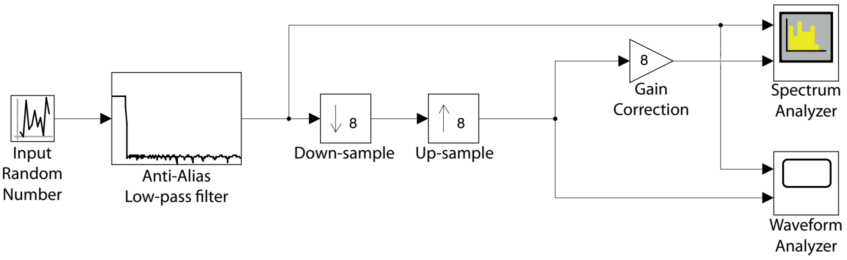
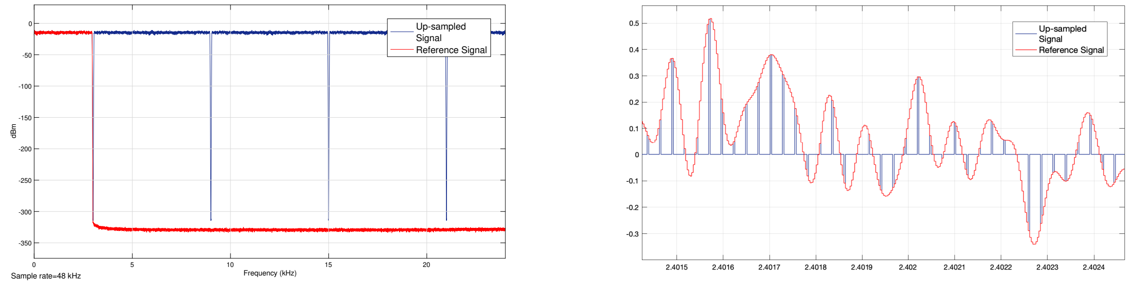


Figure 4.9: Data generator for band-limited signal



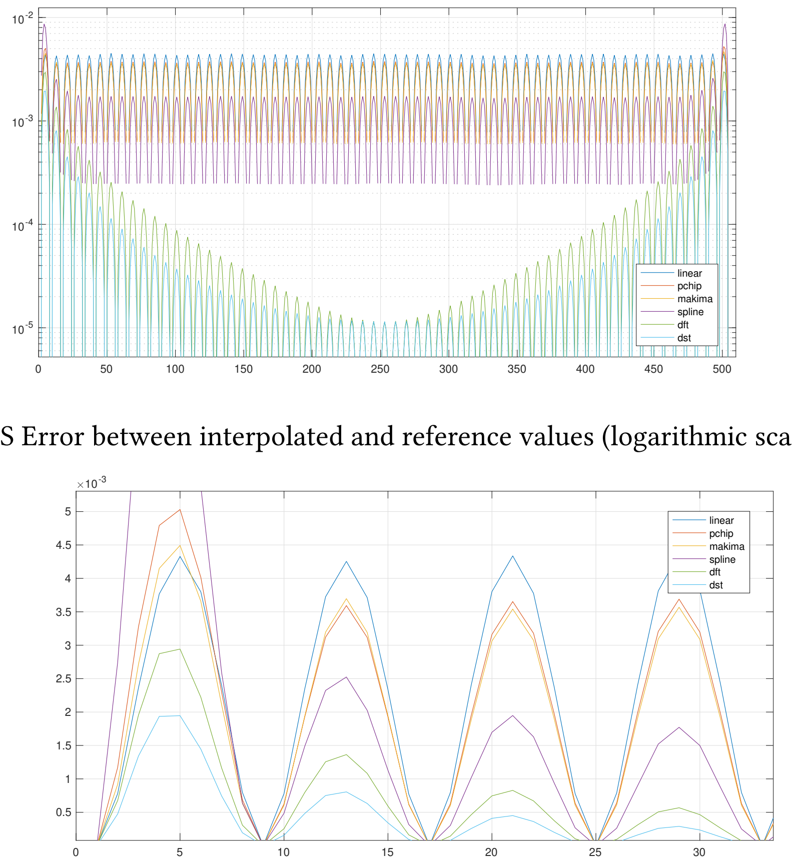
(a) Spectrum of the test signal

(b) Time-domain waveform of test signal

Figure 4.10: Frequency and time analysis of the input signal

The sub-sampled data is interpolated by various methods in Figure 4.11, where the Root-Mean-Square Error (RMSE) between the interpolation and the reference band-limited signal is shown. Data is obtained by averaging the RMSE for 10000 random 64-elements blocks of the input signal. Data is up-sampled 8 times, resorting to a Monte-Carlo analysis. This is performed to even out the signal-related variability. In this analysis different interpolation techniques have been compared: linear interpolation, Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [158], modified Akima piecewise cubic Hermite interpolation (mAkima) [159], spline interpolation [160], DFT interpolation and the proposed enhanced DST interpolation.

All the methods pass through the input points, showing null error values, but between them the error is variable. The worst interpolator is the linear method, followed by the mAkima and PCHIP ones. Then the spline shows good results but the Fourier-based methods offer very good results, in particular in the central region. In the logarithmic plot is possible to see that the proposed method behaves better than the DFT one, while



(a) RMS Error between interpolated and reference values (logarithmic scale)

(b) Particular of RMS Error between interpolated and reference values (linear scale)

Figure 4.11: RMSE comparison for different interpolation methods

both outperform the previous ones at the expense of more computations.

4.2.3 Applications

This interpolation technique is easy to implement and leads to good results also for small signal blocks. Border effects are less evident due to the absence of discontinuities in the interpolation scheme. The original signal is not modified by windowing functions to reduce border artifacts, leading to a better time-domain result. The DST framework is very versatile, it is possible to re-sample the signal with an arbitrary number of points, also a non-integer upsample factor. This is usually difficult to achieve with time-domain interpolation schemes like zero-padding between samples followed by a low-pass filter. Even with an arbitrary number of points, the resulting trigonometric interpolant curve still passes through the original samples, leading to a very good interpolation result. It is also easy to perform a high-quality non-integer re-sampling by varying the ratio of output-to-input points.

2D interpolation

This scheme can be used also for still and moving raster image interpolation. The method is not directly realizable in both dimensions at the same time but it can be performed as a two-passes algorithm. For example it is possible to re-sample in the horizontal direction at first and then in the vertical one. As before, the visual artifacts at the border are very limited. The input image can also be decomposed into smaller sub-blocks to enhance a parallelized computation. It is also possible to rotate the image resorting to the Paeth algorithm [161] as the trigonometric curve can be sampled with sub-pixel accuracy as requested by the shear operation. The main problem with raster images is that they are usually not bandlimited like, for example, audio signals. Abrupt changes of intensity leads to Gibbs oscillations when interpolating, but this is common for all the bandlimited interpolation schemes. It can be solved only by non-linear operations like super-resolution (for example using a Deep Neural Network (DNN)) and other adaptive interpolation methods.

Figures 4.12 and 4.13 show a side-by-side comparison of a sample image and its 4-times enlarged version. In particular, Figure 4.13 details some high-frequency zones that are usually difficult to interpolate correctly.

Numeric integration and derivation

The result of the proposed method is a trigonometric polynomial, a continuous curve that connects the discretized points of the original band-limited time-continuous

(a) Original 512x512 pixels image

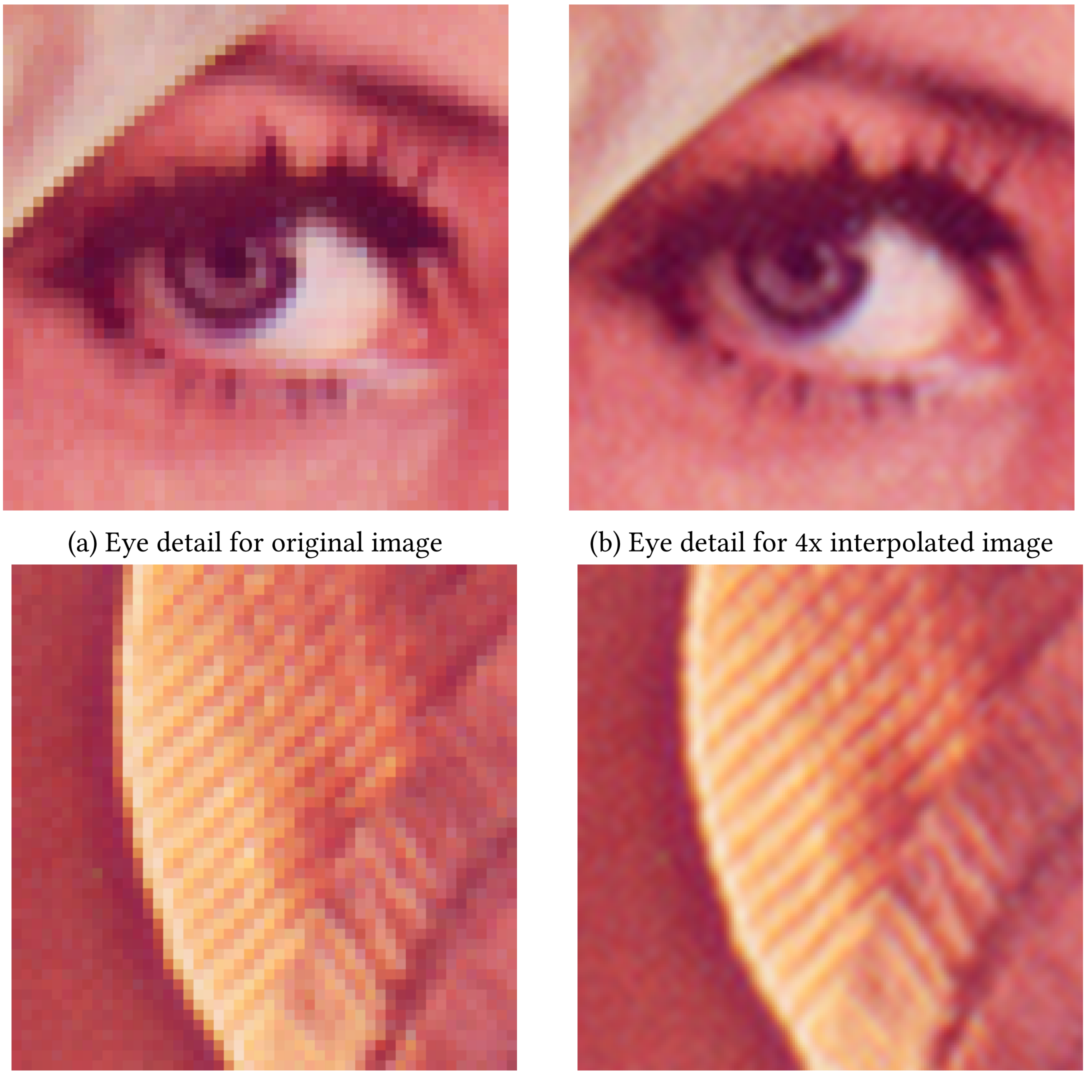
(b) Interpolated 2048x2048 pixels image

Figure 4.12: Image interpolation with the proposed method

signal. Depending on the border conditions, the two curves are usually very similar. This can be exploited to find the integrals and the derivatives (also fractional [162]) of the sampled signal, exploiting the trigonometric integrals and derivatives for the cosine ad the sine terms [163]. In [162] it is easy to notice that the major source of error depends on the derivative dicontinuities at the border when the DCT repetition scheme is used. Contrarily to DCT and DFT-based algorithms, the proposed method can reduce the error at the extrema, leading to more faithful results.

4.2.4 Remarks

This novel interpolation method reduces the border artifacts present in other interpolation schemes, while preserving the reconstruction accuracy. In the middle of long block sizes, the DFT spectral method lead to similar performances but the border parts are affected by the nature of the employed algorithm. The proposed method splits the signal in two oscillatory contributions and applies different interpolation rules stemming from the DCT and the DST approaches. Due to this trigonometric interpolation equivalence, it is possible to analytically manipulate the interpolant curve to extract some useful results. With a multi-pass interpolation, it is also possible to either re-sample or rotate



(a) Eye detail for original image

(b) Eye detail for 4x interpolated image

(c) Hat detail for original image

(d) Hat detail for 4x interpolated image

Figure 4.13: Magnified view of the obtained image

a raster image. As both DCT and DST present good energy-compaction properties, it would be interesting to adapt in the future this algorithm to signal compression tasks. DST and IDST can be computed resorting to fast algorithms, leading to accurate results also in resource and time-restricted scenarios. The real nature of the DST requires less computations than the complex-valued DFT. As the coefficients obtained by the DST are directly the weights of the sinusoidal basis, it is possible to skip the IDST operation and directly re-sample the analytic curve if only a subset of the new points is needed. This can be useful for example in conjunction to a sliding-window approach, similar to the zero-padding and low-pass filter approach used in FIR filter-based interpolation.

Chapter 5

Novel Contributions - Modulation

This chapter deals with issues related to Delta Sigma Modulation. This is a broad topic, well studied with some remarkable results in the state-of-the-art. There are some practical problems to be solved, like the modulator stability and the complexity of look-ahead algorithms. Additionally, some already solved issues can be implemented in more efficient or performant ways.

At first, in Section 5.1, a novel way to create multi-level modulators with integrated Dynamic Element Matching is presented, which achieve analog mismatch error shaping with multiple modulators in parallel.
Then, Section 5.2 presents an elegant way to solve the modulator stability issues by dedicating to the saturation operator, the source of instability, an unconditionally stable noise shaper. This is just a minor modification to the conventional structure but it offers great benefits with low overhead.
Next, Section 5.3 proposes a look-ahead-like structure that exploits time-interleaved modulators to reduce non-linear effects. A delay line matches the group delay of a linear-phase low-pass filter to perform the required time alignment.
Eventually, Section 5.4 offers a low power, low complexity structure based on the Dyadic Digital Pulse Modulation and an analog FIR DAC that can produce a high-quality analog output even with a two-levels bitstream.

5.1 Dynamic element matching as parallel modulators

As explained in Section 3.2, multi-level ΔΣMs need a DEM in order to work correctly in the analog domain. DEM algorithms try to create many two-levels signals that, once summed, return the original multi-level signal, while ensuring that analog errors are shaped outside the audible range. Each algorithm shows different properties depending on its inner working principle. The basic DEM is a pseudo-random scrambling of the available output signals. Despite being simple, it offers no data-directed analog error shaping but only an even element usage, de-correlating the error from the audio input. Using data-directed approaches it is possible to achieve first, second or higher-order static mismatch shaping [105, 164–166].

The majority of DEM algorithms works on the multi-level signal obtained by a previous ΔΣM. Apart from some rare works, like the one in [167], most DEM circuits are implemented outside the main modulator. This means that the DEM has no access to the high-quality signal that was available before the ΔΣM. This severely limits the DEM freedom of choice and it is difficult to embed calibration into this stage, even if this is the part that has direct control over the analog stage.

Under these limitations, it would be interesting to have the DEM directly embedded in the ΔΣ Modulator, with a global control of the sum of the single two-levels outputs. This section aims to provide such a modulator by exploiting dither compensation between output branches.

5.1.1 DEM principle

Dynamic Element Matching (DEM) exploits linear compensation of the available outputs to re-create the multi-level input signal. Every two-levels digital-to-analog stage (DAC element) is affected by a *static gain error* and a *static offset error*, as shown in Figure 5.1. In addition there will be *dynamic errors* mainly due to *Inter-Symbolic Interference (ISI)* and *jitter*, but these will not be considered in this section.

Focusing on a single DAC element, its output can then be written as

$$\mathrm{OUT}_{\mathrm{DAC}} = \alpha \cdot \mathrm{IN}_{\mathrm{DAC}} + \beta \tag{5.1}$$

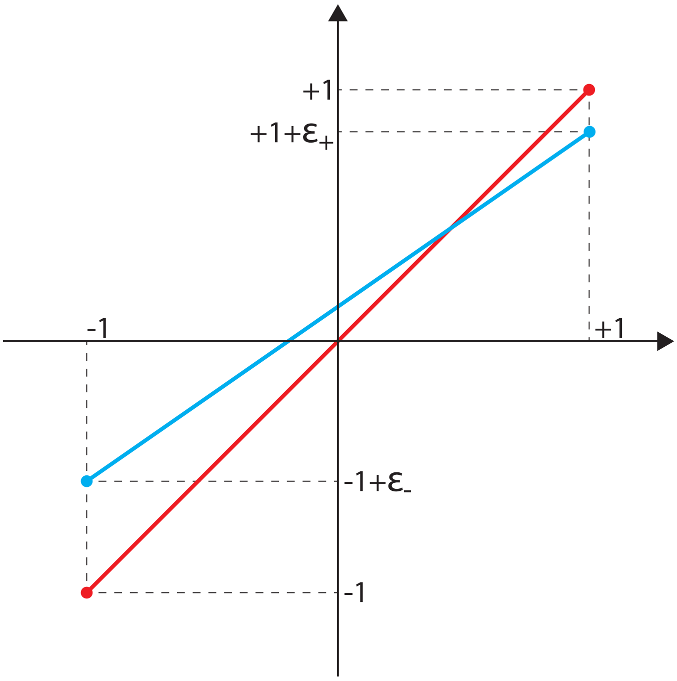


Figure 5.1: A real 1 bit DAC (in blue) shows gain and offset errors compared to the ideal DAC (in red)

where $\alpha = 1 + \frac{\epsilon_+ - \epsilon_-}{2}$ is the gain with error and $\beta = \frac{\epsilon_+ + \epsilon_-}{2}$ is the offset error. Figure 5.1 shows graphically these errors. The line is still straight, as there are only two possible values, so there are no non-linear errors.

A multi-level output is usually composed by the sum of many (possibly weighted) two-levels contributions, as shown in Figure 5.2. As a first approximation, this addition can be treated as perfect beacuse, for small signals, the analog summing part is usually easy to build with a relatively inconsistent error by exploiting the Kirchhoff Current Law (KCL).

Under this assumption it is possible to write

$$\sum^{k} \mathrm{OUT}_{\mathrm{DAC},k} = \sum^{k} \left(\alpha_k \cdot \mathrm{IN}_{\mathrm{DAC},k}\right) + \sum^{k} \beta_k. \tag{5.2}$$

The term $\mathrm{IN}_{\mathrm{DAC},k}$ is formed by the k^{th} two-levels digital signal. The DEM framework exploits time to force the differences between the various α_k to even out to a single gain

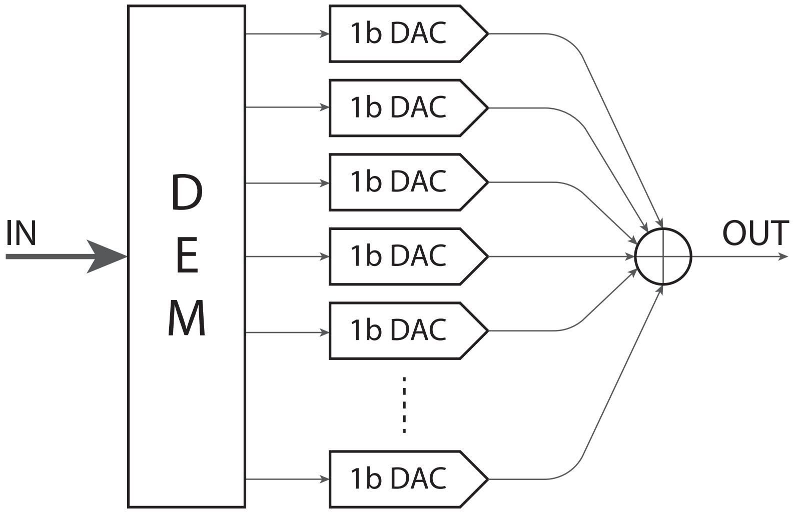


Figure 5.2: Multi-level DACs are composed by many two-levels contributions

coefficient and shaping the remaining error out of band. In other words, DEMs aim to obtain

$$\sum^{k} \mathrm{OUT}_{\mathrm{DAC},k} = \alpha_{\mathrm{TOT}} \cdot \mathrm{IN}_{\mathrm{DAC}} + \mathrm{NTF}(\epsilon) + \sum^{k} \beta_k \tag{5.3}$$

where NTF(ϵ) encompasses the shaped static gain error. Here, the total offset is usually small and negligible. The error is concentrated outside the audio band. It is thus inhaudible and easy to filter out, if necessary. The total gain error is just a small gain error that does not affect the signal frequency response. Equation 5.3 shows a linear input-output relationship, hence the final result will be directly proportional to the input signal.

5.1.2 Parallel Delta Sigma Modulators can perform like DEM

The best way to achieve the DEM goal would be having every two-levels channel carry the audio signal in a constructive way, while the shaped quantization noise in a destructive way. This is a difficult system to build. Classic DEM algorithms, as explained before, are not able to discern the audio input from the quantization noise, and thus they have to split the signal just by following the output of the ΔΣM and knowing the audio band limits to perform the analog error shaping. This severely restricts the DEM freedom of choice and the final output signal quality. On the contrary, using separate ΔΣMs for each DAC channel would ensure that every digital output would contain an equal amount of audio signal (due to ΔΣM signal transfer function). This means, by denoting the shaped error of the $k-th$ element as $\mathrm{NTF}(\epsilon_{q,k})$, that

$$\mathrm{IN}_{\mathrm{DAC},k} = \mathrm{STF}(\mathrm{IN}_{\mathrm{AUDIO}}) + \mathrm{NTF}(\epsilon_{q,k}) \tag{5.4}$$

so that, with a unity signal transfer function,

$$\alpha_k \cdot \mathrm{IN}_{\mathrm{DAC},k} = \alpha_k \cdot \mathrm{IN}_{\mathrm{AUDIO}} + \alpha_k \cdot \mathrm{NTF}(\epsilon_{q,k}). \tag{5.5}$$

By linearity,

$$\begin{aligned}\sum^{k} \left(\alpha_k \cdot \mathrm{IN}_{\mathrm{DAC},k}\right) &= \left(\sum^{k} \alpha_k\right) \cdot \mathrm{IN}_{\mathrm{AUDIO}} + \sum^{k} \left(\alpha_k \cdot \mathrm{NTF}(\epsilon_{q,k})\right) \\ &= \alpha_{\mathrm{TOT}} \cdot \mathrm{IN}_{\mathrm{AUDIO}} + \epsilon_{\mathrm{TOT}},\end{aligned} \tag{5.6}$$

the resulting analog signal from Equation 5.2 will exactly contain a scaled version of the input audio signal and the weighted sum of the quantization errors produced by each independent modulator.

5.1.3 Total error reduction

Equation 5.6 introduced the term ϵ_{TOT}, the total quantization noise given by the combination of the analog gain error and the digital quantization error. As the goal of a ΔΣM is to remove as much error from the signal band as possible, the in-band analog error will be small due to the product between the gain error and the shaped quantization noise.

This means that the total in-band error will be small, even if all the branches have the same output. It is now possible to try to force the output branches to have a destructive interference-like behavior regarding the quantization error ϵ_q. This can be achieved by exploiting the unity STF of each ΔΣM branch, using the scheme presented in Figure 3.4. In particular, it is useful to know that a signal injected at the input of the ΔΣM loop will be superposed at its output to the input audio signal and the remaining quantization noise. Due to the non-linear nature of the ΔΣMs, the modulator will act differently

with different input stimuli. For example, higher magnitude input signals will lower the mean quantization error for two-levels modulators, as the envelope of the signal will be closer to the quantized value. Thus, injecting similar but different signals in two identical ΔΣMs in parallel can potentially lead to different quantization error behaviors. Thanks to the ΔΣM loop, if there are no stability issues, the two errors will be spectrally similar and they will show a constructive and a destructive part between the outputs of the two modulators.
If the injected signal is fed to the two modulators with opposite signs, after the output summation their contribution will even out approximately to zero. At the same time, the input audio signal will present a constructive addition. The result will be a three-level signal containing the input signal plus a reduced quantization noise.

5.1.4 Extension to multiple branches

It is now clear that it is possible to obtain a high-quality multi-level signal stemming from single modulators in parallel. The basic rule for the proposed method is that each modulator receives the input signal plus a support signal that will be canceled by the other branches. If the modulator is stable, N branches will translate to an output of $N + 1$ levels. Under normal modulator operation, the output of the whole system will show a higher SNR than for the single modulator case, both in the digital and the analog domains. The lower recirculating noise enables a further global loop filter to orchestrate the sub-modulators, as shown in Figure 5.3. In this setup, even if the individual branches do not provide a high-SNR individually, the global loop will force the error generated by the entire system to be shaped.

This method can result in an SNR similar to a native multi-level quantizer, but now the DEM mechanism is embedded in the whole system. The DEM-like output is directly compared to the high-quality input digital audio signal, freeing up the system from the usual DEM constraints. The injected signals can be arbitrarily chosen in order to not occupy the audio signal band, so that the delicate audio signal will not be degraded by uncompensated analog errors. These signals can also help with dithering the input audio signal, weakening the ΔΣM-related artifacts.

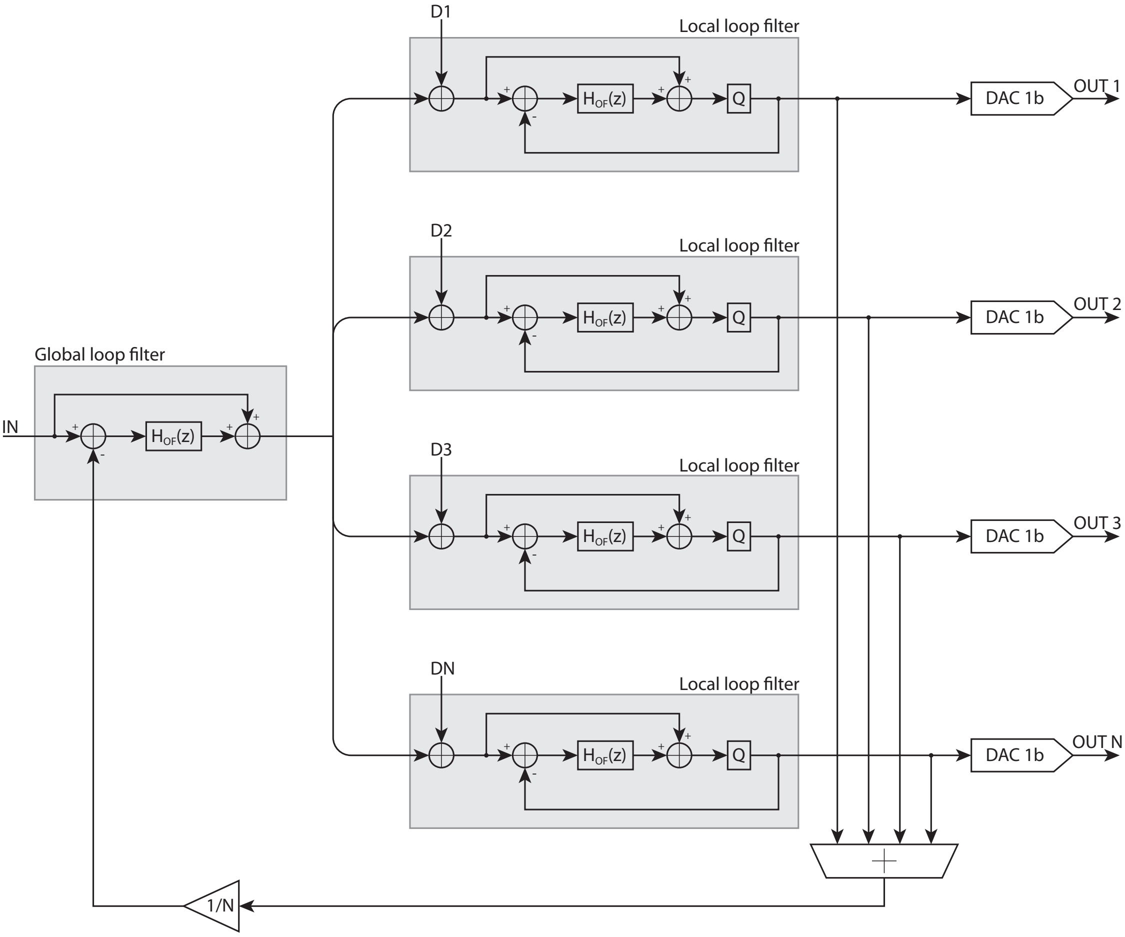


Figure 5.3: Multi-level ΔΣM with parallel branches

5.1.5 Calibration

Now that the DEM is part of the nested feedback loops, it is possible to embed a digital equivalent model for each branch directly in the loops. The single element DAC can be compactly modeled as in explained in Section 3.1.1 as a four entries LUT placed in each branch feedback path to model both static mismatches and ISI.

A small total offset error is usually harmless in audio systems, so this issue can be neglected, but the gain and the ISI error contribution shaping can be crucial to obtain the required output quality. This is true in particular for modern transistor production processes which are very good for digital purposes but problematic for analog applications. If a digital calibration routine is performed correctly, most part of analog errors can be concealed by noise-shaping.

5.1.6 Example

In this section, an example circuit is realized as shown in Figure 5.3. The basic modulator is created stemming from a fourth-order high-pass elliptic filter, following Equation 3.2, with a sampling rate of 6.14 MHz. In this design, the branches and the global modulator have the same loop filter for simplicity. The global feedback loop controls 4 different branches for a total of 5 possible output levels. Each branch has its own auxiliary signal injected after the global loop filter, so that their sum becomes zero. The injected signals are low-amplitude, high-pass filtered pseudo-random noise sources in order to leave the audio band as free of added noise as possible.

Figure 5.4 shows the spectrum obtained by the multi-branch modulator and the spectrum from a single-branch system. The corresponding waveforms are presented in Figure 5.5. The single branch modulator follows the implementation in Figure 5.3 but simply with $N = 1$ and no injected signal. Both spectra present the same noise shaping behavior but in the multi-level case the high-frequency error at the high-end part of the spectrum is greatly reduced. Also the harmonic distortion is lower in the multi-level case. The most important feature in this circuit is not the combined response but the fact that each branch is automatically performing high-order Dynamic Element Matching as explained in Equation 5.6. This is just a simple example to show in practice the beavior of the proposed modulation scheme, but this technique can be improved by adjusting the number of branches, the loop filters (both the global and the branches ones) and the properties of the injected signals. It is important to notice that the injected signal reduces the maximum input amplitude tolerated by the modulator before instability occurs.

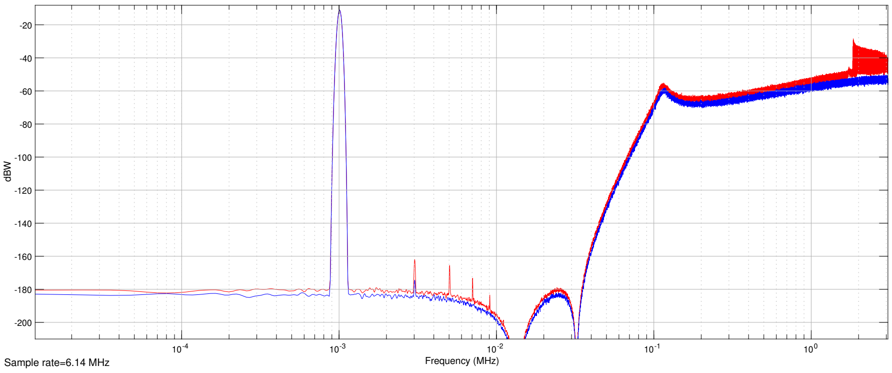


Figure 5.4: Comparison of the output spectrum for a single-branch system (in red) and the four-branches one (in blue)

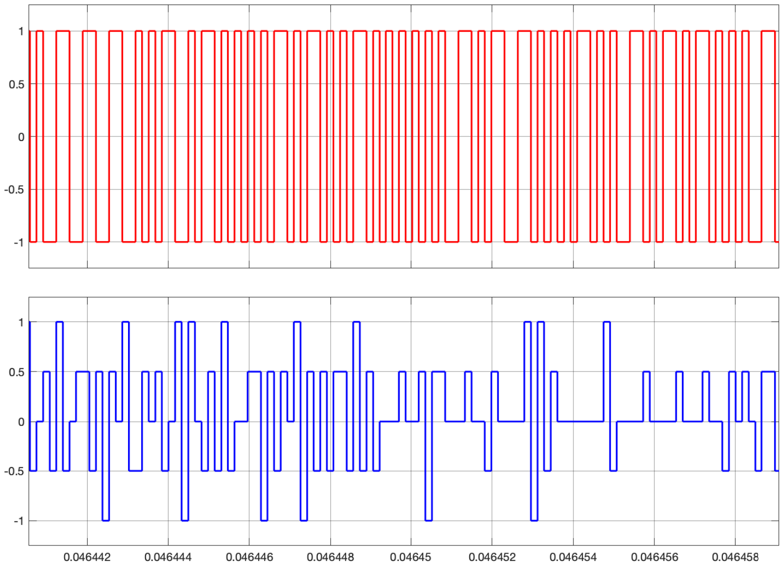


Figure 5.5: Comparison of the output waveform for a single-branch system (in red) and the four-branches one (in blue)

5.1.7 Remarks

The presented structure is able to solve some issues that arise in real-world implementations of a multi-level ΔΣM DAC. It shifts the complexity from an external, inflexible DEM circuit to an easier to design classic ΔΣM. DEM algorithms with comparable analog noise-shaping capability are difficult to design. Each element of the DEM has to be modulated by a ΔΣM (to retain its history and apply noise-shaping) and a Vector Quantizer (VQ) is needed to decide which outputs have to be selected [37, 168]. VQ-based DEMs are difficult to handle and design: due to the feedback structure and limited output capability they tend to show instability issues, like ΔΣMs, but with reduced degrees of freedom to tame this problem. Other simpler matching schemes have problems like limited noise-shaping capabilities, usually first or second-order shaping, but it is possible to realize them efficiently. These can be employed where the analog error is known to be low, so that there is only a small amount of error that remains to be shaped. For modern transistor production processes, the presented method can offer better error shaping, leading to higher SNR values. If also the ISI is shaped, for example by counting and forcing the mean number of transitions to even out or by calibration, also this dynamic error contribution can be shaped using separated branches.

This novel approach to analog error mismatch compensation, however, requires an additional loop filter for each two-levels output. This can be costly to implement, in particular for real-time modulators with many output levels. Easier to implement DEMs with multi-level ΔΣMs can still be a more viable resource when the analog error is limited.

This method can be integrated with conventional ΔΣM design workflow and frameworks, there is no need to learn about VQ-based DEMs. This speeds up the design time and the inherently digital nature of the algorithm makes it compatible with modern mostly-digital transistor technology nodes. Additionally, it is easy to embed a DAC element model in the feedback loops for calibration and to enhance mismatch shaping, and the injected signals can help dither each single branch, thus reducing artifacts. The structure can be straightforwardly expanded to the desired number of levels. Each level will lower the total amount of circulating quantization noise, hence allowing a more aggressive noise-shaping by the global loop filter or, alternatively, a lower OSR. Each branch will have an intrinsic high-SNR and the global loop will enforce noise compensation between branches.

5.2 Increased stability Delta-Sigma modulators

As explained in [38], the first order, two-levels ΔΣM is unconditionally stable. Stability issues arise from higher-order modulators due to quantizer overload. When the signal at the quantizer input has a particular amplitude and time-domain response, the quantizer is unable to return to the loop filter enough energy to make the filter behave correctly. Under these conditions the loop becomes unstable and without stabilization techniques (some of them are discussed in [91]) it is not possible to recover the correct behavior of the modulator.

This section will deal with the stability issue with a simple yet effective method.

5.2.1 Non-linear Delta-Sigma Modulator behavior

Classic ΔΣMs conceptually are just saturating quantizers that exploit oversampling to shape the quantization error spectrum with the aid of feedback and loop filters. The saturating quantizer is the problematic part of the system because the saturation operation makes it impossible to automatically recover the system from instability. this bounded quantizer can be split in two parts, an unbounded quantizer and a saturation element, as shown in Figure 5.6. The quantizer part would be harmless without the saturation element, as overload would never occur in this case. It would just add more quantization error to the system. The saturation part limits the possible number of output levels to match the available levels-budget.

The most common saturating quantizer is the two-levels one. It forces the output signal to take either one of the two available levels. The multi-level quantizers show the same behavior but the saturation mechanism show itself after the n^{th} level. Just as in the two-levels case, this saturation is the source of instability. The idea presented in this work consists on splitting the two non-linearities to aggressively noise shape an unbounded quantization error and then apply the saturation with a mild, highly stable shaper.

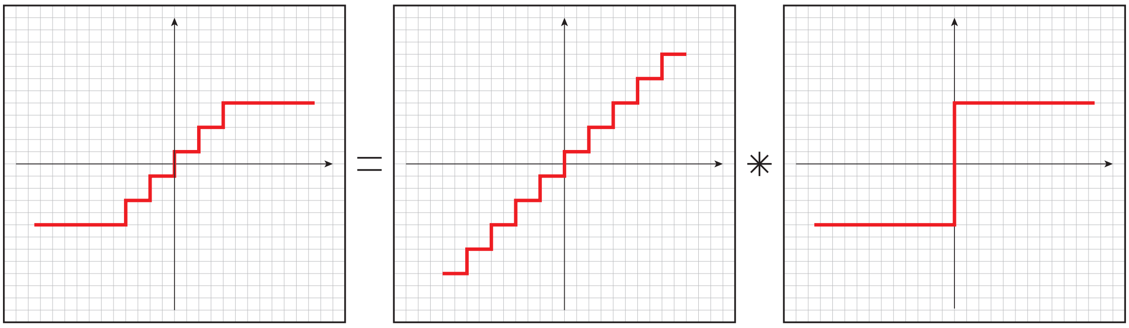


Figure 5.6: A bounded quantizer can be split as an unbounded quantizer followed by a saturation element

5.2.2 Modified quantizer

As the proposed system is focused on some possible digital implementations, it is easy to perform high-accuracy computations and use complex circuits. A modified saturating quantizer is here presented, that will be used to create a highly stable two-levels ΔΣM. The extension to a multi-level quantizer is straightforward and will not be discussed here.

As stated before, the saturation part of the quantizer is the source of instability, so this modified quantizer will perform the quantization, first, and the saturation, then. The simplest way is to use an unbounded multi-level quantizer like in Figure 5.6 but, as the goal is to quantize in the saturation region only, this structure adds an excess of quantization error outside this region. It would be more efficient to avoid excess quantization error build-up. The easiest way is to not produce the quantization error at all in the region outside the saturation limit.

The proposed quantizer is shown in Figure 5.7. It can be seen that the signal is quantized only in the central region, and the transfer function outside it is just a straight line. This translates to a zero quantization error in this region as there is no added quantization. This will help the loop filter to work correctly as no error will be added when the signal exceeds the saturation region. This simplifies the return of the signal inside the saturation region, as will be explained next.

The saturation element, after the main modulator, will then force the output to have only two-levels as specified. In this way, the global behavior will match the original saturating quantizer.

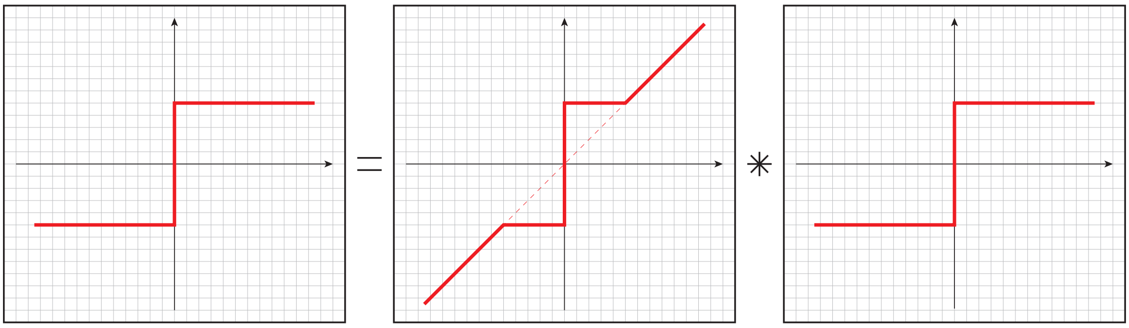


Figure 5.7: Modified bounded quantizer

5.2.3 Modified Delta-Sigma Modulator

Now that is possible to split the saturating quantizer as a quantization operator followed by a saturation operator, a novel stabilization strategy can be applied.
The first stage deals with the quantization error produced in the by the modified unbounded quantizer. It is placed in an high-order modulator. This leads to an unconditionally stable system that will mostly work in the saturation region and will sometimes jump in the unquantized region to ensure stability. This occurs when the sum of the input signal and the shaped noise overflows outside the allowed quantization region. If the loop noise gain is not too aggressive, the overflow event will have a small duration and will be recovered by the filter. Otherwise, in the case of frequent overflows, the efficacy of the presented method would be reduced due to in-band shaped noise build-up. The output of this first part is given as input to the saturation operator. It will create an output word compliant with the number of available elements. The overflow part is then treated as an error source. With an open-loop system, this error would ruin the previous noise-shaping. There is the necessity to re-shape the overflow error to reduce its impact on the final result. The easiest way is to use an unconditionally stable first-order error-feedback noise shaper. In the digital domain this requires a small quantity of computational resources and can be performed exactly.

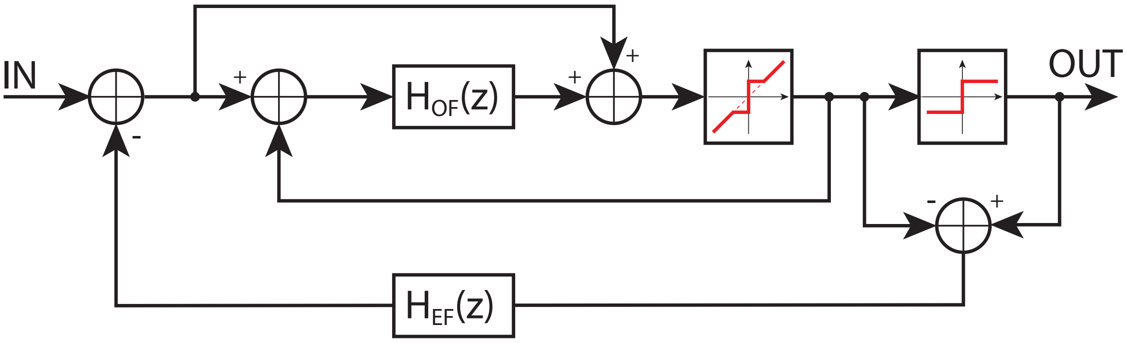
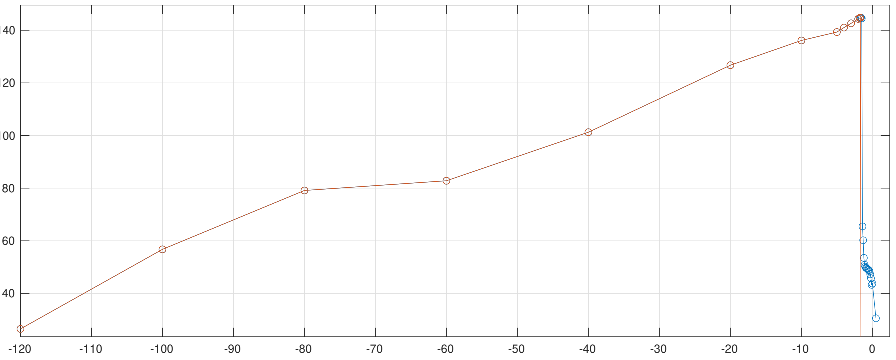


Figure 5.8: Modified bounded quantizer

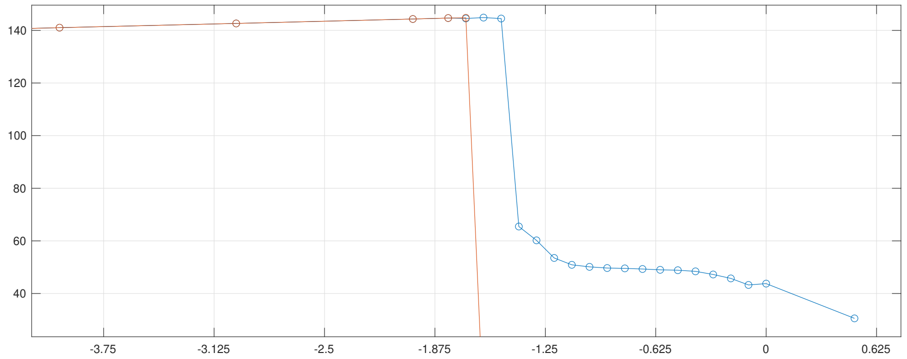
5.2.4 Example

Figure 5.8 shows the result of a possible implementation. As there is no restriction to the first modulator complexity, a high-order modulator is here employed. The saturation element is treated as explained before. Figure 5.9a outlines the obtained SNRs for the traditional ΔΣM structure and for the proposed one. Figure 5.9b details the SNR where the classic modulator reach instability. It is possible to notice that before the instability region both the modulators share the same SNR because the quantizer is not overloaded. At this input level, the saturation part in the proposed scheme does not activate. After the classic modulator fails, the other degrades to a lower SNR due to the added shaped error. The most interesting feature is that this modulator does not become unstable and so it is able to automatically recover to the regular behavior as soon as the input signal amplitude returns to an acceptable value. The classic modulator is not able to resolve the instability issue and it needs a forced reset, or other similar compensation techniques, to return to a stable behavior. Real-life ΔΣMs require a compensation mechanism to correctly work, because hostile input signals are difficult to detect and the stability depends also on the internal state of the loop filter.

It is interesting to notice that the modulator retains its stability also over the 0 dB: the modulator outputs a scaled-down version of the input due to noise, harmonic distortion and the limited output signal energy availability. Nevertheless, it still remains stable and it can return to the high-SNR zone effortlessly. This situation can be very problematic to solve for traditional ΔΣMs.



(a) Audio band SNR for the conventional modulator (in orange) and the proposed modulator (in blue)



(b) Audio band SNR for the conventional modulator (in orange) and the proposed modulator (in blue), which remains stable for the entire input range with a partial output SNR degradation when the classic structure fail

Figure 5.9: SNR of proposed unconditionally stable modulator

5.2.5 Remarks

The proposed approach is a simple yet effective way to automatically ensure Delta-Sigma modulator stability. Contrarily to other algorithms, there is no need to check instability and make decisions based on the output or the integrators' states. This eases the circuit design and no external compensation mechanism is needed, which can be expensive in terms of memory and operations.

This method modulates the noise floor, which can be audible as added noise, but it saves the system from difficult-to-recover instability situations. If the main modulator is well designed, the stabilization mechanism will activate rarely, just when the first stage fails and overflows.

Every real-world ΔΣM requires a stabilization mechanism to work properly while guaranteeing good baseline performances (when the modulator is stable). Previous stabilization attempts were focused on recovering the circuit from instability by constraining the ΔΣM internal states and sometimes use expensive state-reverting techniques to return to a stable behavior. This work, instead, lets the modulator work freely using a modified unbounded quantization operator which has no stability issues. In the proposed example, the noise overflow issues are corrected automatically by a highly stable first-order error-feedback structure. No state-detection circuitry is needed and the main modulator performance is not directly affected by saturation. This is a remarkably simple solution to obtain stable high-order modulators at the expense of a steep drop in SNR after the overload. It is mainly focused on a digital implementation due to the delicate error feedback structure. In future works it would be interesting to try to employ higher order modulators with low noise gain also in the saturation section.

5.3 Time-interleaved streams

In Digital to Analog Converters, many issues arise from multi-level architectures due to relative errors between output elements, hence the need to employ Dynamic Element Matching algorithms. This is a limitation that is not present if there is a single element, a 1 bit output. Compared to multi-level architectures, 1 bit Delta-Sigma Modulators are nevertheless the worst-case scenario in the Delta Sigma framework [15] as they feature:

- the highest quantization error,
- emphasized non-linear behaviors (spurious idle tones, harmonic distortions etc.),
- difficult proper dithering due to quantizer saturation,
- quantizer gain and linearity highly dependant on input signal statistics and loop filter structure,
- higher risk of instability (in particular with aggressive noise-shaping),
- higher out-of-band noise (which has to be filtered out),
- less linearity, hence less matching between the simplified linear analysis and effective behavior.

Some of these issues can be partially mitigated or solved by look-ahead techniques but the computational workload is very high and thus it is not easy to implement them in real-time. It would be interesting to have a reliable and fast way to create high-quality bitstreams.

This section proposes a novel way to time-interleave bitstreams to create high-quality output.

A similar solution is presented in [169] that can further enhance the proposed technique.

5.3.1 Return-To-Zero coding

Return-to-Zero (RTZ) coding is a technique developed in the telecommunication field. As stated by its name, this technique fills half the duration of a digital signal with a logic zero value, a sort of "rest position". While it requires a doubling in symbol frequency over the simpler Non-Return-to-Zero (NRZ), it ensures that there is a zero-valued part for each produced symbol. In the frequency domain, this is a doubling of sampling frequency with a symmetrical signal spectrum repetition in the newly available frequency region. This means that the RTZ signal spectrum can be straightforwardly derived from a classic NRZ of half the sampling frequency by zero-stuffing it in the time domain, a factor of two upsampling, which generates the related spectral image.

Two RTZ signals with a half-symbol delay can thus be linearly summed together to create a new two-levels signal in virtue of their alternate zero-valued stuffing. They can be thought of as two time-interleaved RTZ signals forming a NRZ signal with a doubled sampling frequency.

This principle can be extended to more signals by extending the number of zero-valued samples stuffed between two adjacent samples.

Due to zero stuffing, the energy carried in the original spectrum section by each stream is reduced by $N + 1$ times the amount of added zeros. This issue can be taken into account when exploiting time-domain interleaving.

5.3.2 Stream interleaving

As stated before, some classic 1 bit ΔΣM issues can be solved by look-ahead techniques. They are prohibitively expensive for a real-time application. Their working principle is to have the modulator look in the future stream of input samples for the best combination of output symbols to match them under some criteria, or cost metrics. An example fot this is to try all the possible outputs to find the combination that leads to the minimum in-band error while retaining stability. The problem is that the computational workload grows exponentially as m^n times the non-look-ahead scheme, where m is the number of output levels (or possible symbols) and n is the number of look-ahead samples. Due to the exponential nature and the fact that $m \geq 2$, the amount of total computations grows fast. In [26] some heuristics are presented to reduce the required efforts but the look-ahead techniques still remain restricted to an off-line pre-computation.

The main limitation in conventional ΔΣM systems is that the loop filter acts on a sample-by-sample basis. It cannot predict the error that will be created by future samples. It will try to correct the error on-the-go and it will go out of stability if unable to do so, due to quantizer saturation. Some works try to implement a real-time simplified look-ahead structure using a modified quantizer consisting of Look-Up Tables [170, 171] but they are limited to only a few look-ahead samples and require huge tables.

In this work, interleaving is exploited to partially "look-ahead in the future". At first, a two-levels signal is created by a classic ΔΣM. It modulates the input signal inside the output stream, embedding it in shaped quantization noise. It creates also non-linear errors like harmonic distortion and spurious tones. In particular, for the audio spectrum, these errors are the main audible artifacts. It would then be interesting to correct them somehow.

Here, this is performed thanks to a combination of a linear-phase low-pass filter and a second auxiliary modulator, as depicted in Figure 5.10. At first, the input audio signal is subtracted from the stream produced by the first modulator, the top grey block, and the result is fed to a low-pass filter, the bottom red block. The filter, whose structure dictates most of the required hardware, should be an FIR of Type II to introduce a half sample delay. It is mandatory to time-align the outputs of the two modulators for this RTZ time-interleaving approach. The resulting filtered signal will match the in-band error produced by the first modulator plus some unattenuated high-frequency content, depending on the filter quality. Due to the linear-phase approach, the error produced by the first modulator will be detected before its effective generation by the first modulator, depending on the filter length. This is the reason why this method can be considered as a sort of "look-ahead technique", the error starts showing in advance when compared to classic ΔΣM approaches. The first output will need a simple delay line to match the group delay of the linear-phase filter.

Now that the in-band error produced by the first modulator is isolated, it can be inverted and re-modulated by the second ΔΣM with unitary gain STF, the bottom grey block, to apply a linear compensation mechanism. If this second modulator has to deal only with noise and small spurious tones, this dither-like signal will excite the non-linear part of the modulator. It will mainly add harmonic distortion of a noise-like error source, which is just other attenuated noise. Due to the exact digital summation of the two signals, the second one will correct the in-band error created by the first modulator exactly, up to the low-pass filter quality. The total in-band error will be approximatively equal to

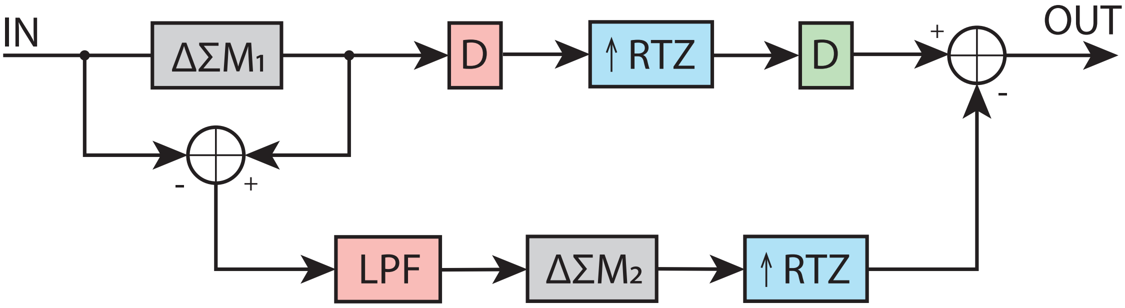


Figure 5.10: Time-interleaved modulators block scheme

the one produced by the second modulator alone. The OOB error will approximately consist on the sum of the ones produced by both the modulators. If they are statistically uncorrelated noises, their amplitude will average out.

The output of the main modulator has to be time-aligned with the second stream using a delay element (top red block) to match the group delay due to the linear-phase low-pass filter. Both the streams are upsampled by a factor of two and interleaved (blue and green blocks).

The modulators can have different structures: the first one, which carries the whole input signal, can be realized with a shallow noise shaper just to guarantee a relatively low in-band noise for the second modulator. A more aggressive shaper can be used to lower the in-band noise floor in the second ΔΣM.

The resulting signal transfer function, for the upsampled input signal, is simply

$$\mathrm{STF} = \frac{1}{2} z^{-2\mathrm{GD}-1}, \tag{5.7}$$

where GD is the low pass filter group delay, and the noise transfer function is

$$\mathrm{NTF_{TOT}} = \frac{1}{2} \left[\mathrm{NTF}_{\Delta\Sigma\mathrm{M}_1} \left(z^{-2\mathrm{GD}} - H_{\mathrm{LPF}} \right) - \mathrm{NTF}_{\Delta\Sigma\mathrm{M}_2} \right]. \tag{5.8}$$

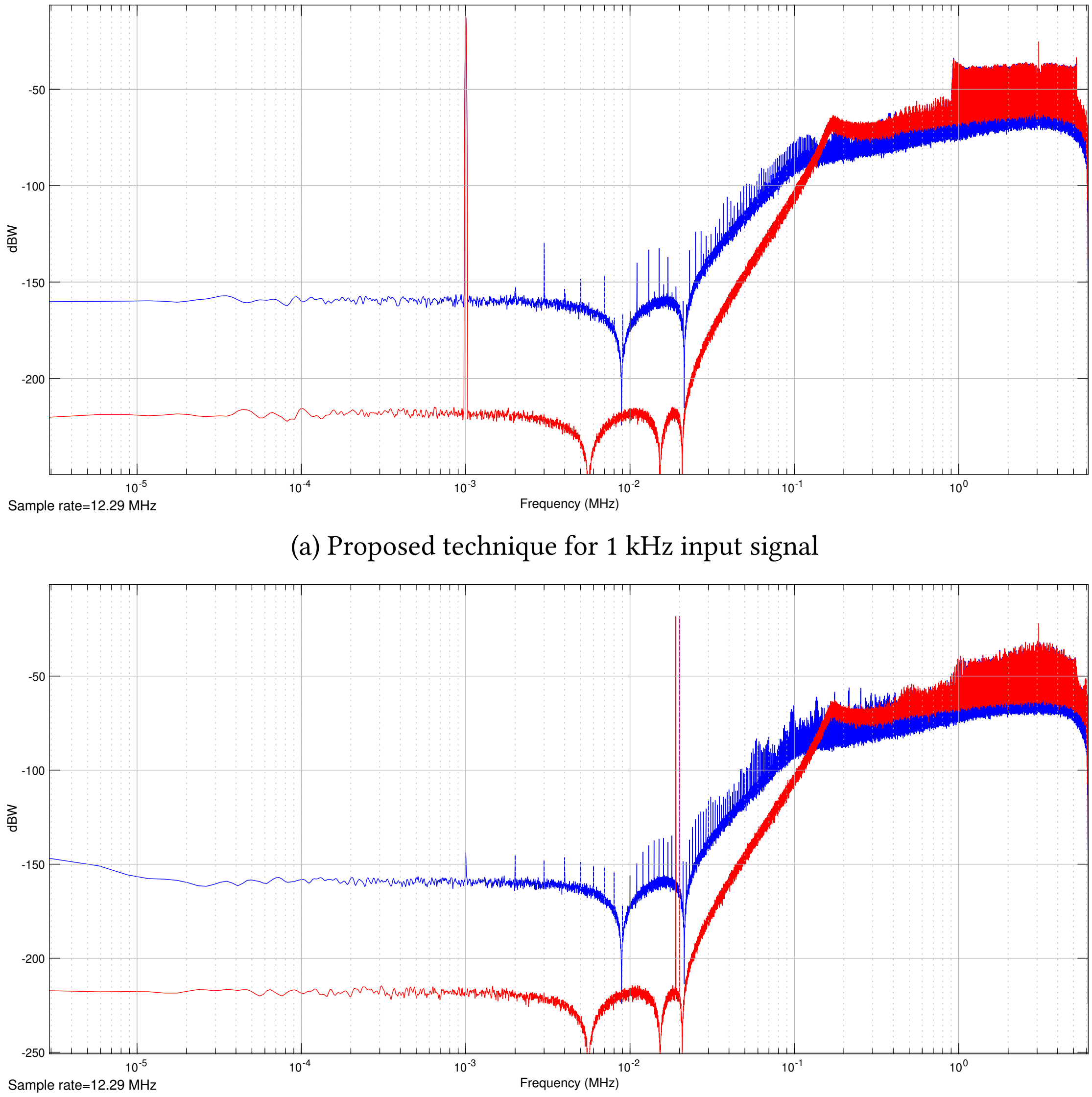
5.3.3 Example

Two ΔΣMs NTFs have been designed, stemming from two different elliptic filters, to test the proposed architecture. The first one features a limited noise shaping to enhance the maximum signal amplitude before instability. The second one shows an aggressive noise shaping with extended transition band. The employed low-pass filter is an equiripple-designed FIR filter with a length 25 samples, with linear phase and a group delay of 12.5 samples. The time-alignment part is thus 12 samples long and the half-sample delay is introduced after the RTZ upsampling to time-interleave the two signals.
Figure 5.11 show the first modulator spectrum in blue and the total output in red. It is possible to notice that in the baseband the system shows a very high linearity, even if no dither is applied. There are no visible harmonic distortion nor inter-modulation tones.

5.3.4 Remarks

The presented work proposes a workaround to the complexity of look-ahead modulators. It exploits two time-interleaved streams in which the second compensate the in-band error produced by the first one. Look-ahead modulation is provided thanks to a linear-phase filter, whose time-domain error detection, the "look-ahead" part, depends on the length of the filter. This translates to excellent time-domain signal reconstruction, with vanishing low in-band artifacts and non-linear behaviors. This is due to the second modulator, which has to deal with a combination of noise, small harmonic distortion components, and some small spurious tones. All these artifacts are already greatly attenuated by the first ΔΣM.

For multi-bit signals it is easy to apply the same technique of error isolation and compensation. Usually, multi-level signals have already good time-domain performances due to the lower quantization error circulating in the system so this method will be less effective. The presented technique is similar to the Multistage Noise-shaping (MASH) approach [38, 172], but with the substantial differences that the second modulator works only on the in-band residual error. Additionally, there is no necessity of output filters as



(a) Proposed technique for 1 kHz input signal

(b) Proposed technique for 19 and 20 kHz input signals for IMD behavior

Figure 5.11: Time-interleaved output spectra for 1 kHz and IMD test signals. In blue the output of the simple modulator, in red the total output of the time-interleaved stream

the second stage is assumed having unitary gain STF, like in the Sturdy MASH (SMASH) approach [173].

Unfortunately, only one of the two modulators to carry the whole input signal and this limits the maximum total output signal to half the maximum level of the first modulator. This can be too restrictive but it is possible to partially insert the input signal in the second modulator, if stability is not affected. Thanks to a constructive sum, the total audio signal carried by the final bitstream will be boosted by the second modulator. The drawback is that the second modulator will create its own non-linear artifacts related to the input signal, like harmonic distortion, lowering the effectiveness of the presented method.

5.4 DDPM-based DAC

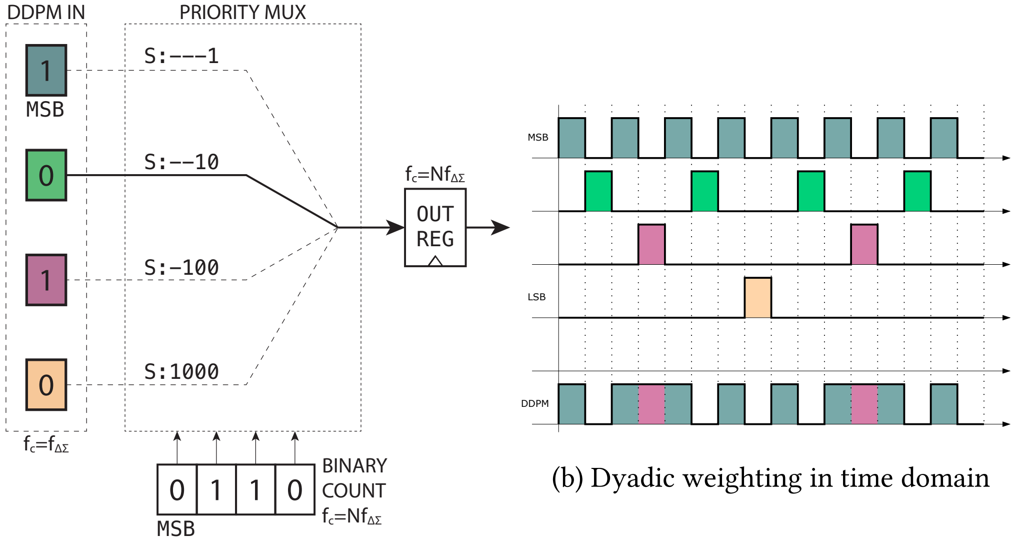
The Dyadic Digital Pulse Modulation (DDPM) is a simple yet effective way to trade quantization in amplitude for quantization in time [118]. Contrarily to the Pulse-Width Modulation (PWM) scheme, which presents a similar high frequency behavior, the DDPM does not introduce harmonic distortion of the input signal but only modulation-related noise. This is a rather important property when designing a modulator for audio, as the linearity is one of the most important features needed by audio applications. In this section, the DDPM is exploited after a multi-level ΔΣM to obtain a 1 bit output signal, and an FIR-DAC structure [174] is used to mitigate the OOB noise created by the DDPM itself. The result is a multi-level signal with intrinsic DEM properties, as the signal is, in first approximation, equally carried by all the output elements. The shift-register-like DAC setup acts as a moving-average digital filter, whose frequency response is well known to follow the sine cardinal (*sinc*) function like the Cascade of Integrator-Comb (CIC) filter [133].

5.4.1 DDPM review

DDPM is a process that trades a $(2^N - 1)$-to-2 level number reduction in amplitude to a 2^N increase in clock frequency, e.g. a seven-levels converter at frequency *Fs* becomes a two-levels stream at *8Fs*. The Dyadic nature of the algorithm creates a non-uniformly distributed spectrum, whose peaks are mainly concentrated at integer multiples of the baseline sampling frequency *Fs*. Similar behavior is obtained using Pulse-Width Modulation (PWM) but DDPM offers superior in-band performances. Plain PWM adds harmonic distortion and the first frequency spike occurs at the original signal sampling frequency. This generates audible spurious tones generated by Intermodulation Distortion (IMD) and other non-linear effects [175]. On the contrary, DDPM behaves, on a first approximation, like a Zero-Order Hold (ZOH) in the original signal band. It introduces only a minor error source at frequencies near the original sampling one. The ZOH-like frequency response can be pre-compensated for maximum performances.
A key DDPM factor relies in its intrinsic simplicity: it can be realized as a priority multiplexer and a binary counter with a very little computational effort, as shown in

Figure 5.12. This implementation is even simpler than the plain PWM one as no algebraic part is needed, reducing the signal propagation critical path and enabling very high theoretical working speeds. The priority multiplexer can be realized with a chain of simple two-ways multiplexers or with some additional combinational logic, depending on speed requirements. The binary counter can be built efficiently as a simple asynchronous counter requiring only N flip-flops.

The main drawback is related to Inter-Symbolic Interference (ISI). Contrarily to PWM, which fixes the mean number of signal transitions independently of the input signal statistics, the DDPM number of transitions heavily depends on the input signal. This can be problematic, in particular for newer technology nodes where the analog properties of transistors are neglected to offer better digital performances. Fortunately, on a first approximation, this error is deterministic and can be addressed by digital calibration methods, as proposed in [118]. For example, the DDPM ISI error source can be modeled with a Look-Up Table (LUT) inside a multi-level ΔΣM before the DDPM amplitude-to-time conversion.



(a) DDPM simplified block scheme

(b) Dyadic weighting in time domain

Figure 5.12: DDPM structure

5.4.2 FIR-DAC

An analog Finite Impulse Response (FIR) DAC is the analog equivalent of a conventional digital FIR filter. It can be built as a chain of digital flip-flop registers, whose outputs are linearly weighted using either transistors, resistors of capacitors. This mathematically translates to a Direct Form structure whose algebraic description is

$$y[n] = B_0 x[n] + B_1 x[n-1] + \cdots + B_N x[n-N] = \sum_{i=0}^{N-1} B_i \cdot x[n-i], \tag{5.9}$$

where $x[n]$ are the samples input to the filter at time n, $y[n]$ is the filter output, $\boldsymbol{B_i}$ are the multiplicative coefficients that determine the filter response and N is the order of the filter.

The simplest FIR-DAC features equally-weighted multiplicative coefficients, i.e. all the $\boldsymbol{B_i}$ are equal to one. This case is well studied and presents a deterministic response. This is called Moving Average Filter (MAF). It behaves like a poor quality linear-phase low-pass filter, with an amplitude response in the frequency domain that follows the *sinc* function. Equal weights are particularly interesting as all the analog elements can be designed as nominally equal.

In the digital domain, where errors stem only from quantization-related errors, this filter can be implemented exactly.

In the analog domain there will always be some mismatch-related errors that will divert the system from its ideal behavior. The ideal MAF cannot thus be realized as an analog circuit. If the input signal can have more than two-levels the delay elements in the filter have to be implemented as expensive and possibly unreliable Sample-And-Hold (SH) circuits or, alternatively, as a digital shift register with individual multi-level DACs. If the input signal is limited to two-levels, the delay element can be easily implemented as 1 bit DACs resorting to a digital shift-register and a two-levels DAC per delay element. This structure is easy to manufacture and integrate on modern Complementary Metal–Oxide–Semiconductor (CMOS) technological nodes. It is a reliable structure with rather stable performances. Additionally, the MAF FIR-DAC intrinsically works as a DEM circuit. Each output element carries the entire signal (time-shifted by the delay elements) plus some shaped quantization noise, as explained in Section 5.1.2. The time-shifting requires a pre-compensation scheme to make sure that the output signal has the correct amplitude, due to the *sinc*-like frequency response of this filter. This pre-emphasis of high frequencies is not difficult to embed in previous filtering stages if needed. Otherwise, due to the oversampling nature of the ΔΣMs, the introduced high-frequency drop can be inaudible by the final listener. The compensation can be safely neglected in this case.

5.4.3 Delta-Sigma modulator

The DDPM could work on a Nyquist-rate signal but the high word-width would require a very fast operational frequency. For example for a 16 bit signal the required DDPM frequency would be in excess of 3GHz, which is difficult to achieve and the dynamic analog errors would easily ruin the signal. Also, the DDPM self-noise introduced near its first spectral peak could be detrimental to output quality.

This issue can be addressed by a multi-level noise shaper with oversampling. The resulting ΔΣM can work at moderately high frequencies to reduce the DDPM self-noise effect in the audio band. The quantizer word-width should not be too high, to ensure a bearable switching frequency after the DDPM. The ΔΣM maximum frequency value will depend on the analog circuitry performances, as dynamic errors usually worsen with the increase of switching frequency due to parasitic reactive elements.

For example, a good candidate could be a ΔΣM at a working frequency of about 768 kHz and a 5 bits output with the uppermost level unused (31 levels in total), which will produce a DDPM bitstream at less than 25 MHz.

This modulator can be designed with conventional techniques. As always, it is recommended to use a unity STF to retain the original signal unmodified and leave the complex interpolation and pre-compensation filters outside the delicate ΔΣM loop. As stated before, it is possible to encompass a LUT in the loop to model the DDPM ISI error and shape it at this stage. In this case, a calibration technique is required to successfully model the ISI contribution.

5.4.4 Example

To validate the proposed method, two similar designs have been implemented. Both the modulators feature 31 levels at 768 kHz with TPDF dither spanning two levels. The DDPM translates this to a two-levels bitstream at 24.576 MHz. An equally-weighted FIR-DAC with 32 elements is used to attenuate the spectral peaks produced by the DDPM. Figure 5.13 shows schematically the signal flow diagram.

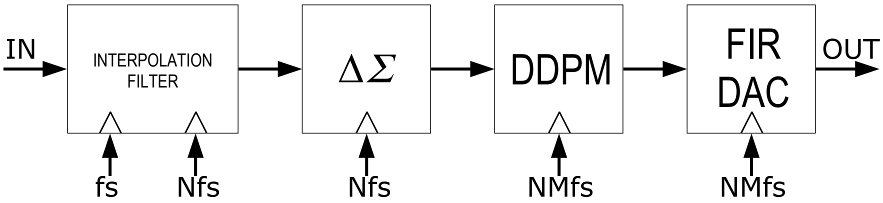


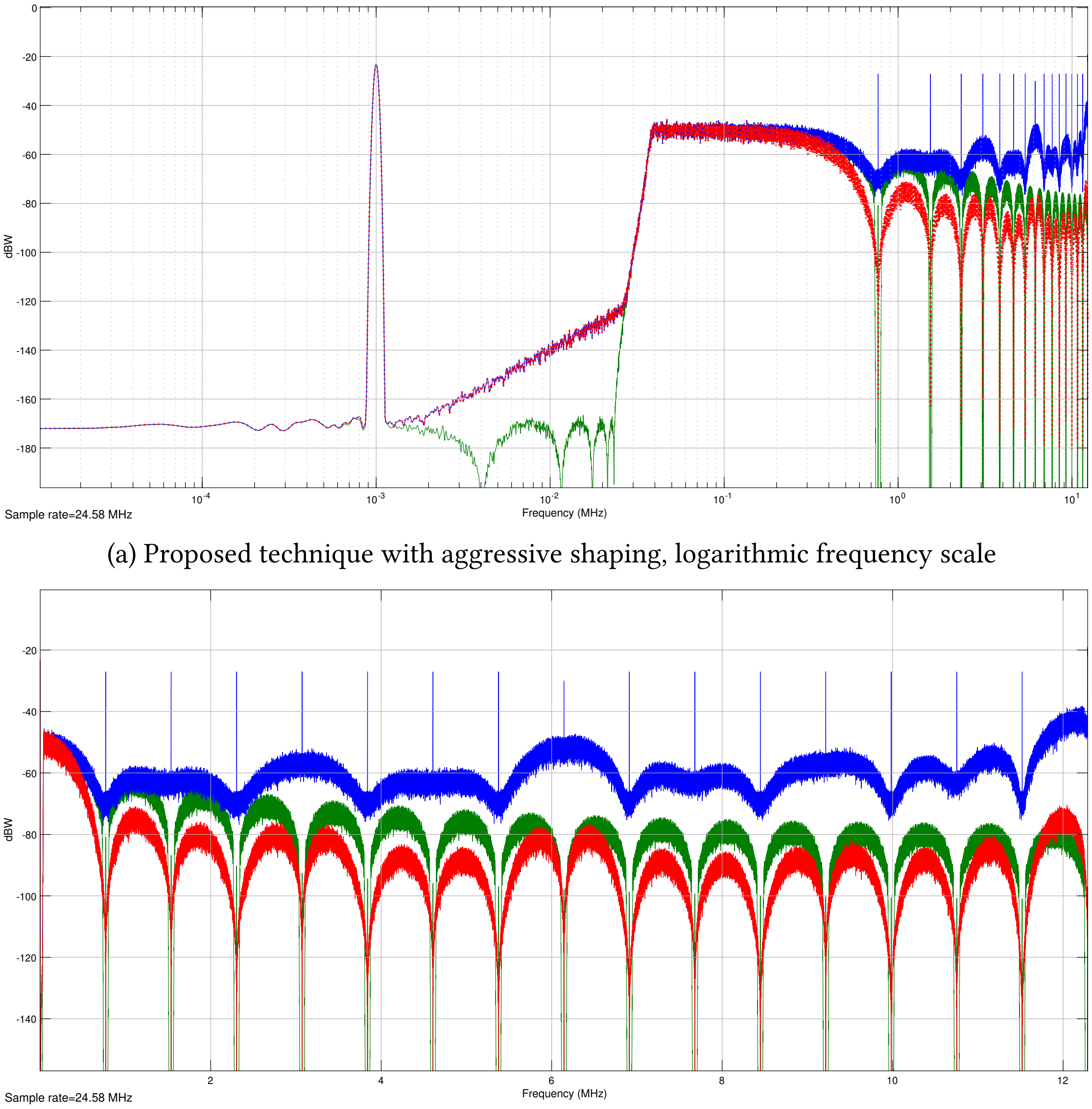
Figure 5.13: Proposed system simplified block scheme

The first solution features an aggressive noise shaping. This limits the maximum input signal amplitude before instability. Figure 5.14 shows the obtained spectra. The output of the ΔΣM, ZOH-upsampled, is the green spectrum. Thanks to the TPDF dither its behavior is perfectly linear. The output of the DDPM is the blue spectrum. As predicted, it produces multiple peaks at integer multiples of the ΔΣM Nyquist frequency. These peaks are compensated using the FIR-DAC.
Overall this is a good low-power modulator but it is easy to notice that the DDPM artifacts extend also in the audio band, in particular in the high-frequency zone.

It is thus possible to design a less aggressive shaper to take this phenomenon into account. This solution allows a higher maximum input signal amplitude before instability.

Custom Prototype

Thanks to the elegance and simplicity of the proposed solution, a simple prototype Printed Circuit Board (PCB) has been designed to assess real-world performances. The



(a) Proposed technique with aggressive shaping, logarithmic frequency scale

(b) Proposed technique with aggressive shaping, linear frequency scale

Figure 5.14: Proposed technique with aggressive shaping. The DDPM self-noise reduces in-band SNR, in particular at high frequencies

board goal is to create a high-quality mixed-signal (digital and analog) stage to bypass the output capabilities of a common Field-Programmable Gate Array (FPGA). This board takes as input two independent channels, each made of 8 distinct equally-weighted elements, for a total of 16 two-levels DACs. Each DAC present a high-speed CMOS D-type flip-flop with balanced output transistors to ensure a high-quality switching

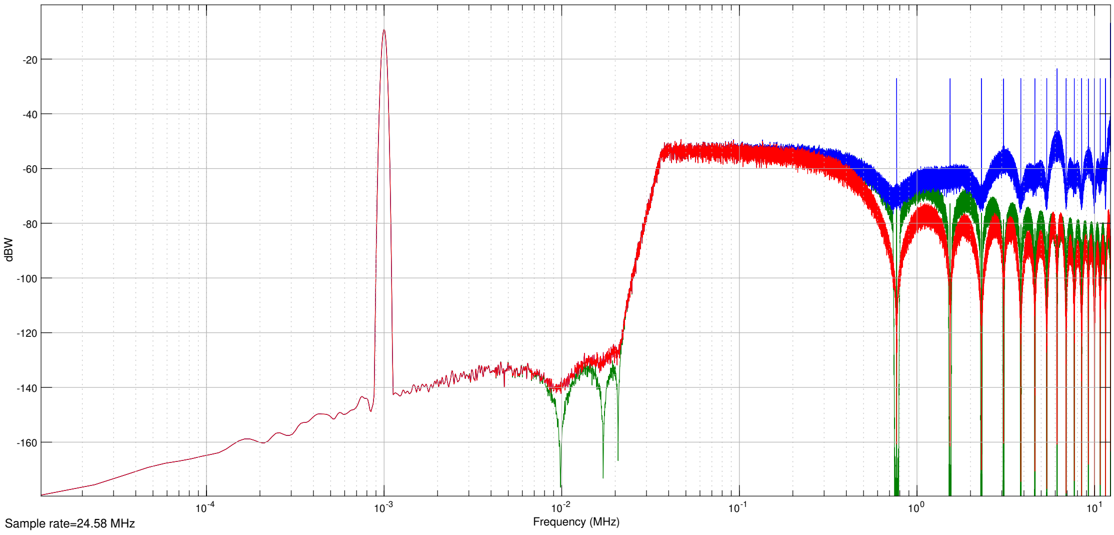


Figure 5.15: Proposed technique with reduced error shaping. The DDPM self-noise matches the ΔΣM shaping capability

stage. This is performed by two dedicated octal SOIC-20 FF by Texas Instruments, the SN74LVT574. The chips are clocked by the FPGA and they shares the same clock signal. After the memory elements, eight matched 100 kΩ resistors in a SOIC-16 package are placed per channel (Vishay's NOMCT16031003AT1). They act as voltage-to-current elements. Next, DIP switches are used to select which elements can be used to test different configurations. The outputs of the switches are routed to a node where the current contributions are summed. An operational amplifier fixes this node to the FF mean output voltage, namely $V_M = \left|\frac{V_{HI}+V_{LO}}{2}\right|$. This makes the input behave as a bipolar signal, avoiding DC offsets and balancing the voltage excursion of each element. This is done by a dual precision ultra-low-noise operational amplifier, the AD8599 from Analog Devices, which acts both as a transimpedance amplifier to convert the current to voltage and as a low-pass reconstruction filter to further reduce the signal OOB noise. The amplifiers outputs can be used as stand-alone audio outputs (with a two-channel 3.5mm audio jack) or can be fed to a precision instrumentation amplifier, the AD8429 from Analog Devices, to reconstruct a single-ended signal if the two channels are configured in the FPGA to be used differentially. This signal can be measured using a coaxial BNC connector. This allows reducing the common-mode noise stemming from preceding stages. The ICs have dedicated ultra-low noise linear regulators, the LT3042 for the

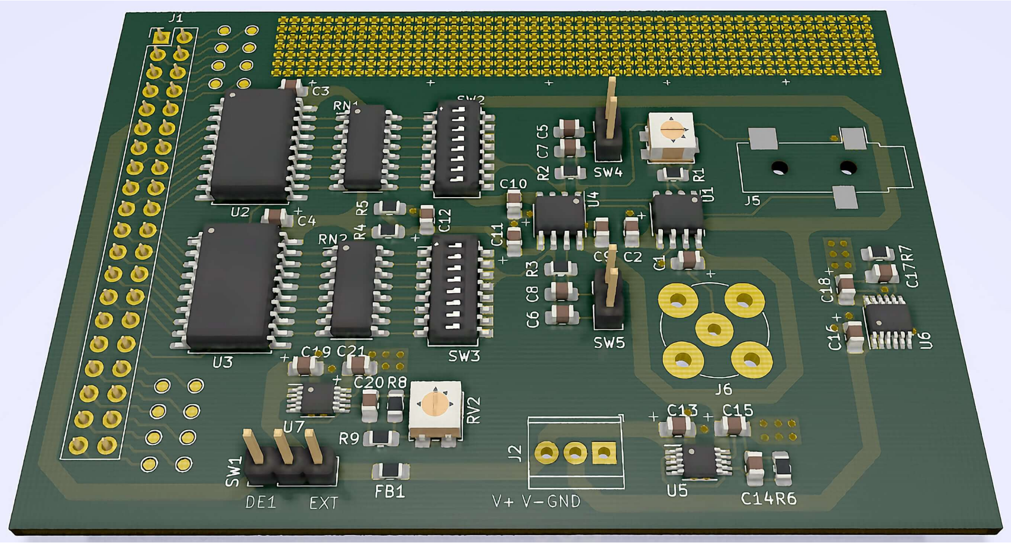


Figure 5.16: 3-D render of the PCB designed to test the proposed system

positive supply and the LT3093 for the negative supply, both from Analog Devices. The PCB is designed to be pin-compatible with the Altera DE1-SoC by Terasic [176]. This development board features a Cyclone V System-on-Chip (SoC) which host a dual-core ARM A9 hard processor. The FPGA fabric can be programmed using an Hardware Description Language (HDL) like VHDL or Verilog. The FPGA implements the I2S receiver port, the multi-level ΔΣM, the DDPM, and the output shift register, which is directly connected to the external FFs in the custom-designed PCB.

Figure 5.16 shows a 3-D render of the board. The left part hosts the connector to the FPGA board. The upper part is filled with a rectangular matrix of unconnected vias for general purpose prototyping purposes. The rear of the board is filled with a solid ground plane. U2 and U3 are the input FFs. RN1 and RN2 are the precision resistor networks. U4 is the dual operational amplifier IC. U1 is the instrumentation amplifier. U5 and U7 are the positive regulators and U6 is the negative one.

5.4.5 Remarks

The proposed technique can exploit the advantages of the DDPM while retaining its key properties without the needing very high clock frequencies. Its usage, in conjunction with a dedicated pre-processing stage, a multi-level Delta-Sigma modulator, and the

shift register-like analog FIR DAC shows promising results.
DDPM is a very effective way to trade amplitude quantization for time quantization but it has to work on a limited number of bits, as the output frequency grows exponentially with respect to the number of input bits. The ΔΣM is the ideal candidate for this task and multi-level, low-frequency modulators show higher linearity and stability versus a similar two-levels one working at the DDPM frequency. With the extended time budget (the amount of time available for each operation) it is possible to apply low-power digital design techniques to the ΔΣM implementation and scheduled resource utilization to lower the silicon area utilization.

At this design point, the analog circuitry has been implemented in a custom-designed board but it would be interesting to design a dedicated CMOS IC with dedicated digital and analog circuits. The digital part is rather simple to implement and the Dynamic Element Matching property ensures high-SNR output analog signal. It is easy to extend the output to a differential scheme just by inverting the main output bitstream, with very low computational effort (an inverter gate and some FFs). The analog DAC elements can be implemented as current-steering DACs, which are compatible with high-speed differential outputs, or as switched-capacitor circuits, which are less affected by ISI and jitter errors and can be easily integrated on silicon with small relative mismatch error.

Chapter 6

Conclusion

Oversampling Delta-Sigma Modulators form a broad topic that can be exploited in various fields, even if this thesis focused only on the audio one. The novel techniques presented in **??** are directed to mostly-digital implementations to fully exploit modern CMOS digital capabilities, even in the presence of poorly-linear analog circuitry.

The proposed ideas are often orthogonal to each other in the solution space and could potentially be applied altogether to build a very high-quality Digital to Analog Converter. Some solutions can be enhanced by embedding a digital model of the analog stage in a closed-loop feedback structure, but this calibration part has not been addressed in this thesis. It would be interesting to work on this feature to help the state-of-the-art advance further.

The purpose of this work was to tackle the most algorithmically complex modern DACs digital building blocks one by one, looking for alternative ways to efficiently implement them or increase the resulting quality. From the initial interpolator to the final Dynamic Element Matching, every aspect of the system has been approached by custom solutions that can be built with off-the-shelf electronic components, like FPGAs, or manufactured into custom modern silicon technologies. Following this document, with a background in digital design and signal processing, it is possible to create a mostly-digital audio DAC with state-of-the-art performances in multiple ways.

Two different approaches have been presented for the interpolation issue, one in the time domain and one in the frequency domain. Both can offer significant advantages in the field and the second one can be useful also for image and video processing.

Contrarily to traditional design techniques, with the DEM outside the main ΔΣM loop, the solution presented in this thesis offers a different perspective to achieve multi-level performances that were previously available only in conventional techniques. The complexity is shifted to the main Delta-Sigma loop, so the computational power could become a bottleneck due to the feedback structure. On the other hand, it soften the constraints associated with classic DEM algorithms, allowing more performant analog error mismatch shapers.

These structures could anyway be expensive to manufacture and utilize, due to the relatively high amount of computations required. They are intended to be used in high-end stand-alone audio DACs, where the price point and the power consumption can be as high as needed for ultimate performances. For casual audio listeners, the DDPM-based proposal can be a competitive solution due to its high output quality, even at low power and silicon area requirements. This algorithm can be interesting for modern True Wireless Stereo (TWS) earphones, which features a very tight power budget and small size, or for the smartphone market segment, where customers demand high-quality audio outputs.

The stability issues of ΔΣMs can be addressed with the proposed dedicated approach, by splitting the bounded quantizer working principle. This workaround effectively solves one of the major problems in the field with an elegant solution. No instability detection techniques have to be implemented, only a minor modification to the original Delta-Sigma structure is required.

Overall, there are many other enhancements to the field that could be investigated, like the use of outphasing signals for the time-interleaving technique or deeper exploitation of the PWM signal features (like the fixed mean number of transitions for ISI error shaping and the deterministic position of frequency peaks) but they are left for ongoing future works.

Appendix A

Other works

Other published works, unrelated to the topic of this thesis, will be briefly listed here. They were produced during the doctoral programme on secondary topics like custom processor design, video coding, artificial intelligence and low power techniques.

Live Demonstration: Tactile Events from Off-The-Shelf Sensors in a Robotic Skin [177]

The humanoid robot iCub needed a bio-inspired, human-like tactile sensor. This issue was addressed by a matrix of capacitive sensors for each finger. A custom multi-core RISC processor with resource sharing was designed for real-time control of the commercial sensors. The processor has been successfully implemented on an FPGA for a live event demonstration.

Approximate-Computing Architectures for Motion Estimation in HEVC [178, 179]

This work aims to reduce the power consumption of the Motion Estimation (ME) block in a High Efficiency Video Coding (HEVC) encoder by exploiting the approximate computing paradigm. After analyzing the ME adders, multiple approximate alternatives are proposed. Results show a tradeoff between power saving and mean ME error.

An Optimized Partial-Distortion-Elimination-Based Sum-of-Absolute-Differences Architecture for High-Efficiency-Video-Coding [180, 181]

To address the high power consumption of ME, a custom low-power hardware accelerator is presented here. It employs algorithmic and technology-dependent optimizations, resulting in fast, low-power and low-energy circuits. Partial Distortion Elimination (PDE) cleverly skips unnecessary computations. Clock gating disables the register switching activity, lowering the power consumption. The structure is synthesized for a low leakage process to further increase the circuit efficiency.

Edge computing: A survey on the hardware requirements in the internet of things world [182]

The Internet of Things (IoT) is a recent trend towards a fully-connected society. Billions of devices will have access to the Internet in future years. It is possible to exploit this feature to decentralize the computing capabilities of a cloud infrastructure. Edge computing consists on using them to pre-process or post-process data locally. Unfortunately, these systems often show power limitations, so it is mandatory to exploit ultra-low power designs to solve this problem. This work reviews the state-of-the-art for various Edge computing aspects.

Low-Power Hardware Accelerator for Sparse Matrix Convolution in Deep Neural Network [183]

Deep Neural Networks (DNN) are flourishing in these years, but they require an enormous amount of computations. Using techniques like the ReLU nonlinearity, many tensor elements drop to zero, leading to a sparse matrix. If this situation is preemptively detected, the related computation can be avoided. Here, an hardware accelerator for DNN convolution is presented. It exploits sparsity to optimize power consumption and execution time.

VLSI Architectures for the Steerable-Discrete-Cosine-Transform (SDCT) [184, 185]

Modern video compression schemes heavily employ the 2-D DCT for its energy compaction property, leading to high compression ratios. Recently, a directional 2-D DCT called Steerable-DCT has been presented. It can compress even further the incoming data at the expense of additional computations. It is thus important to accelearate this execution block by dedicated hardware. This paper propose an optimized solution to cope with this issue.

Appendix B

Digital Serial Interfaces

This appendix aims to give a basic understanding of the two most used consumer digital audio interfaces, namely the I2S and the S/PDIF (with AES/EBU). As explained in the introduction, this is not a comprehensive list and more interfaces are getting implemented in high-end DACs. Serial interfaces are required to move digital data from the digital source to the DAC using a reduced number of wires compared to full parallel decoding. The drawback is that the operating frequency is higher than a parallel interface and then a *deserializer* circuit must be employed. The operating frequencies usually are not prohibitive due to the reduced bandwidth of audio signals. For example, DXD data with 24 bit at 384 kHz in I2S for two channels require less than 20 MHz of clock frequency, which is fairly easy to implement in modern CMOS designs.

B.1 Inter-IC Sound Interface (I2S)

The I2S is a synchronous interface designed by Philips in 1986. It needs one data line (*serial data*, SD) and two clock lines (*serial clock*, SCK, and *word select*, WS). This interface is a very basic one and it is the easiest to build and handle. The SD channel sends Single Data Rate (SDR) Non-Return-to-Zero (NRZ) push-pull serial data clocked by SCK. The WS channel, synchronized to the falling edge of SCK, indicates if the data refers to the right channel (logic 1) or the left channel (logic 0). The working principle is shown in Figure B.1. As it was originally intended for inter-chip communication, there is no standard interconnecting cable and each manufacturer offers different possibilities,

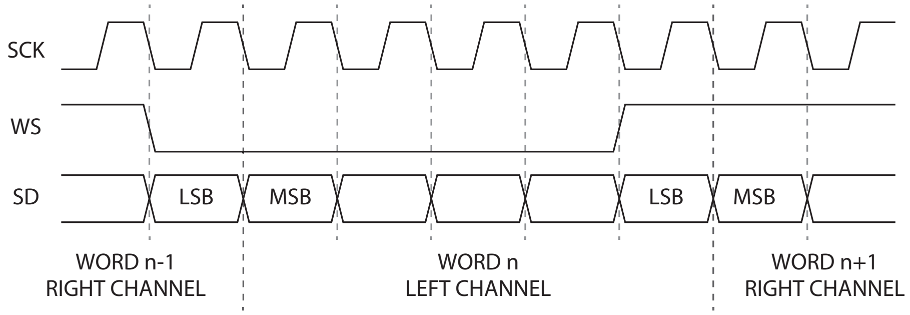


Figure B.1: I2S three-wire protocol

ranging from BNC connectors, to 8P8C connectors (usually RJ45, the same connector as the Ethernet protocol), D-sub connector, HDMI connector, DIN connector, multiple RCA connectors and so on. Some of them are more robust to electromagnetic interference and are better suited to send I2S over longer distances.
I2S has no error correction mechanisms and no negotiation between the sender and the receiver. Data is sent as two's complement signed number with the Most Significant Bit (MSB) first, left-justified. Data starts from the MSB on the clock cycle after the WS transition.

B.2 Sony/Philips Digital Interface Format (S/PDIF)

S/PDIF is a single data wire asynchronous standard, invented by Sony and Philips in 1985. It is the consumer-oriented derivation of the professional-oriented AES3 (or AES/EBU) standard. These two interfaces are very similar, there are only some minor protocol differences but they employ different connection cables. S/PDIF mainly uses unbalanced coaxial 75Ω cables with RCA or optical fiber (the Toshiba Link (TOSLINK)) connections, and a lower voltage than AES3, which limits the maximum distance to about 10 meters. AES3 also uses coaxial 75Ω cables for the unbalanced version with a BNC connector. In conjunction with a higher voltage level it makes up to 100 meters transmission possible. AES3 balanced employs shielded twisted pair 110Ω cables with the XLR connector and a higher voltage, reaching up to 1000 meters transmissions. Both the standards use Biphase Mark Code (BMC) to send both data and clock on a single

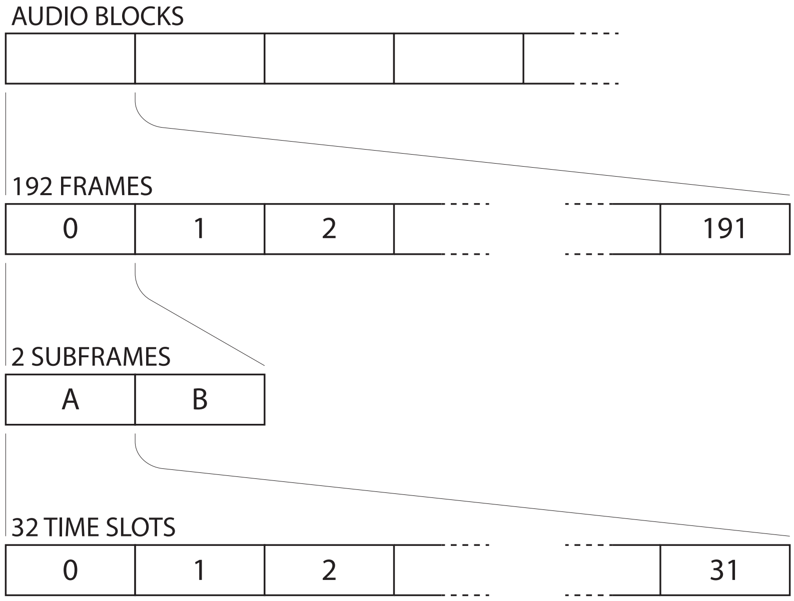


Figure B.2: S/PDIF single wire protocol

wire by encoding the data on a 1-0 or 0-1 sequence to transmit a logic 1 and a 1-1 or a 0-0 sequence to transmit a logic 0, so that there is a polarity inversion at least once in every two cycles. BMC force a constant average value, which eases the transmission and clock reconstruction. This interface is more difficult to integrate as it is not as straightforward as I2S due to the self-clocking scheme and the more complicated protocol structure, schematized in Figure B.2. Data is sent in *audio blocks* each made by 192 frames. A frame is composed of two subframes (one for each channel in a stereo configuration) composed by 32 BMC time slots that contain both PCM data and status information. There is also a preamble used for synchronization and subframe recognition and one channel status value for each subframe, leading to 192 status values per audio frame. Each bit represents different information about the data source and a final Cyclic Redundancy Check (CRC) to validate the received status. For high-quality audio applications, the incoming clock should be discarded once the input frequency has been retrieved by internally handling the signal as a mathematical entity, disjoint from the physical clock. This technique is called Asynchronous Reclock (AR). It avoids using the clock embedded in the S/PDIF stream, which is in practice unreliable due to jitter.

Bibliography

[1] Alan Palmer. “How the Ear Works and Why Loud Sounds Cause Hearing Loss.” In: *Audio Engineering Society Conference: UK 18th Conference: Live Sound*. Audio Engineering Society. 2003.

[2] Marina Bosi and Richard E Goldberg. *Introduction to digital audio coding and standards*. Vol. 721. Springer Science & Business Media, 2012.

[3] Francis Rumsey. “Hear, Hear! Psychoacoustics and Subjective Evaluation.” In: *Journal of the Audio Engineering Society* 59.10 (2011), pp. 758–763.

[4] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT press, 2011.

[5] Claude Elwood Shannon. “Communication in the presence of noise.” In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[6] International Electrotechnical Commission et al. *Audio recording–Compact disc digital audio system*. Tech. rep. Technical Report IEC 60908, 1999.

[7] Eng Tan and B Vermuelen. “Digital audio tape for data storage.” In: *IEEE spectrum* 26.10 (1989), pp. 34–38.

[8] Peter Bloomfield. *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[9] James Boyk. *There’s Life Above 20 Kilohertz! A Survey of Musical Instrument Spectra to 102.4 KHz*. URL: https://www.cco.caltech.edu/~boyk/spectra/spectra.htm.

[10] BM Oliver, JR Pierce, and Claude E Shannon. “The philosophy of PCM.” In: *Proceedings of the IRE* 36.11 (1948), pp. 1324–1331.

[11] Alec H Reeves. “The past, present and future of PCM.” In: *IEEE Spectrum* 2.5 (1965), pp. 58–62.

[12] M Vest. “The advantages of DXD for SACD.” In: *Resolution Magazine* (2004).

[13] Jan Verbakel et al. "Super Audio CD Format." In: *Audio Engineering Society Convention 104*. May 1998. URL: http://www.aes.org/e-lib/browse.cfm?elib=8475.

[14] Peter Nuijten and Derk Reefman. "Why Direct Stream Digital (DSD) is the best choice as a digital audio format." In: *Audio Engineering Society Convention 110*. Audio Engineering Society. 2001.

[15] John Vanderkooy and Stanley Lipshitz. "Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications." In: *Audio Engineering Society Convention 110*. May 2001. URL: http://www.aes.org/e-lib/browse.cfm?elib=9903.

[16] Peter Thorpe et al. "DSD-Wide. A Practical Implementation for Professional Audio." In: *Audio Engineering Society Convention 110*. Audio Engineering Society. 2001.

[17] Joshua D Reiss. "A meta-analysis of high resolution audio perceptual evaluation." In: *Journal of the Audio Engineering Society* (2016).

[18] Julian Dunn. "Anti-Alias and Anti-Image Filtering: The Benefits of 96-kHz Sampling Rate Formats for Those Who Cannot Hear Above 20 kHz." In: *Audio Engineering Society Convention 104*. Audio Engineering Society. 1998.

[19] Karlheinz Brandenburg and Gerhard Stoll. "ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio." In: *Journal of the Audio Engineering Society* 42.10 (1994), pp. 780–792.

[20] Marina Bosi et al. "ISO/IEC MPEG-2 advanced audio coding." In: *Journal of the Audio engineering society* 45.10 (1997), pp. 789–814.

[21] Karlheinz Brandenburg. "MP3 and AAC Explained." In: *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Sept. 1999. URL: http://www.aes.org/e-lib/browse.cfm?elib=8079.

[22] J. Robert Stuart and Peter Craven. "A Hierarchical Approach to Archiving and Distribution." In: *Audio Engineering Society Convention 137*. Oct. 2014. URL: http://www.aes.org/e-lib/browse.cfm?elib=17501.

[23] Philips Semiconductors. *I2S bus specification*. URL: https://web.archive.org/web/20070102004400/http://www.nxp.com/acrobat_download/various/I2SBUS.pdf.

[24] Robert A. Finger. "AES3-1992: The Revised Two-Channel Digital Audio Interface." In: *J. Audio Eng. Soc* 40.3 (1992), pp. 107–116. URL: http://www.aes.org/e-lib/browse.cfm?elib=7056.

[25] Lars Jonsson. “Ebu standardisation of audio over ip for contribution systems.” In: *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. Audio Engineering Society. 2011.

[26] Erwin Janssen and Arthur van Roermund. *Look-Ahead Based Sigma-Delta Modulation*. 1st. Springer Publishing Company, Incorporated, 2011. ISBN: 940071386X.

[27] J. Robert Stuart and Peter G. Craven. “The gentle art of dithering.” In: *Journal of the Audio Engineering Society* 67.5 (May 2019), pp. 278–299. DOI: https://doi.org/10.17743/jaes.2019.0011.

[28] Xue-Mei Gong. “An Efficient Second-Order Dynamic Element Matching Technique for a 120-dB Multi-Bit Delta-Sigma DAC.” In: *Audio Engineering Society Convention 108*. Feb. 2000. URL: http://www.aes.org/e-lib/browse.cfm?elib=9214.

[29] J. W. Bruce and Peter Stubberud. “Circuit Switching Topologies for Dynamic Element Matching Data Converters.” In: *Audio Engineering Society Convention 105*. Sept. 1998. URL: http://www.aes.org/e-lib/browse.cfm?elib=8407.

[30] Ivar Løkken, Trond Sæther, and Anders Vinje. “Segmented Dynamic Element Matching Using Delta-Sigma Modulation.” In: *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*. June 2007. URL: http://www.aes.org/e-lib/browse.cfm?elib=13973.

[31] A. Sanyal, L. Chen, and N. Sun. “Dynamic Element Matching With Signal-Independent Element Transition Rates for Multibit $\Delta\Sigma$ Modulators.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 62.5 (2015), pp. 1325–1334.

[32] C. Noeske, M. Ortmanns, and Y. Manoli. “A dynamic-Element-Matching architecture using individual element error shaping.” In: *2008 51st Midwest Symposium on Circuits and Systems*. 2008, pp. 462–465.

[33] H. T. Jensen and J. F. Jensen. “A low-complexity dynamic element matching technique for reduced-distortion digital-to-analog conversion.” In: *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*. Vol. 2. 1999, 1–4 vol.2.

[34] A. Sanyal and N. Sun. “Dynamic Element Matching Techniques for Static and Dynamic Errors in Continuous-Time Multi-Bit $\Delta\Sigma$ Modulators.” In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 5.4 (2015), pp. 598–611.

[35] R. Schreier and B. Zhang. “Noise-shaped multibit D/A convertor employing unit elements.” In: *Electronics Letters* 31.20 (1995), pp. 1712–1713.

[36] N. Sun. "High-Order Mismatch-Shaped Segmented Multibit ΔΣ DACs With Arbitrary Unit Weights." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 59.2 (2012), pp. 295–304.

[37] Alexander Lavzin, Mucahit Kozak, and Eby G Friedman. "A higher-order mismatch-shaping method for multi-bit Sigma-Delta Modulators." In: *2008 IEEE International SOC Conference*. IEEE. 2008, pp. 267–270.

[38] S. Pavan, R. Schreier, and G. C. Temes. "High-Order Delta-Sigma Modulators." In: *Understanding Delta-Sigma Data Converters*. 2017, pp. 83–116.

[39] C. W. Farrow. "A continuously variable digital delay element." In: *1988., IEEE International Symposium on Circuits and Systems*. 1988, 2641–2645 vol.3.

[40] LR Rabiner et al. "Some comparisons between FIR and IIR digital filters." In: *Bell System Technical Journal* 53.2 (1974), pp. 305–331.

[41] S. Sreekanth, P. B. Shinde, and G. V. Durga. "Performance Analysis of Higher Order FIR Polyphase Filter." In: *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2018, pp. 669–672.

[42] R. M. Deshmukh and R. Keote. "Design of polyphase FIR filter using bypass feed direct multiplier." In: *2015 International Conference on Communications and Signal Processing (ICCSP)*. 2015, pp. 1640–1643.

[43] N. B. Ameur et al. "FPGA implementation of polyphase decomposed FIR filters for interpolation used in Δ-Σ audio DAC." In: *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*. 2009, pp. 1–4.

[44] X. C. Xiong Chenghuan et al. "Design and implementation of a high-speed programmable polyphase FIR filter." In: *ASIC, 2003. Proceedings. 5th International Conference on*. Vol. 2. 2003, 783–787 Vol.2.

[45] Y. J. Yu, D. Shi, and R. Bregovic. "On the complexity reduction of polyphase linear phase FIR filters with symmetric coefficient implementation." In: *2009 IEEE International Symposium on Circuits and Systems*. 2009, pp. 277–280.

[46] Chao Wu, Wei-Ping Zhu, and M. N. S. Swamy. "Design of Mth-band FIR filters based on generalized polyphase structure." In: *2006 IEEE International Symposium on Circuits and Systems*. 2006, 4 pp.

[47] H. Johansson and O. Gustafsson. "Linear-phase FIR interpolation, decimation, and mth-band filters utilizing the farrow structure." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 52.10 (2005), pp. 2197–2207.

[48] S. K. Mitra, A. Mahalonobis, and T. Saramaki. "A generalized structural subband decomposition of FIR filters and its application in efficient FIR filter design and implementation." In: *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* 40.6 (1993), pp. 363–374.

[49] M. D. Lutovac and L. D. Milic. "Approximate linear phase multiplierless IIR halfband filter." In: *IEEE Signal Processing Letters* 7.3 (2000), pp. 52–53.

[50] A. N. Willson. "Desensitized halfband interpolation filters." In: *2007 50th Midwest Symposium on Circuits and Systems*. 2007, pp. 1034–1037.

[51] X. Zhang. "Design of Mth-band FIR linear phase filters." In: *2014 19th International Conference on Digital Signal Processing*. 2014, pp. 7–11.

[52] S. Muramatsu and H. Kiya. "An extended overlap-add method and -save method for sampling rate conversion." In: *1994 IEEE International Symposium on Circuits and Systems (ISCAS)*. Vol. 2. 1994, 313–316 vol.2.

[53] S. Muramatsu and H. Kiya. "Extended overlap-add and -save methods for multirate signal processing." In: *IEEE Transactions on Signal Processing* 45.9 (1997), pp. 2376–2380.

[54] Xiaoxia Zou, S. Muramatsu, and H. Kiya. "The generalized overlap-add and overlap-save methods using discrete sine and cosine transforms for FIR filtering." In: *Proceedings of Third International Conference on Signal Processing (ICSP'96)*. Vol. 1. 1996, 91–94 vol.1.

[55] J. D. Kene. "Extended overlap-save and overlap-add convolution algorithms for real signal." In: *2007 IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007)*. 2007, pp. 539–541.

[56] J. G. Kuk, S. Y. Kim, and N. I. Cho. "An overlap save algorithm for block convolution with reduced complexity." In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 605–608.

[57] M. J. Narasimha. "Modified Overlap-Add and Overlap-Save Convolution Algorithms for Real Signals." In: *IEEE Signal Processing Letters* 13.11 (2006), pp. 669–671.

[58] L. S. Resende, C. A. F. Rocha, and M. G. Bellanger. "A linearly-constrained approach to the interpolated FIR filtering problem." In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. Vol. 1. 2000, 392–395 vol.1.

[59] R. Lyons. "Turbocharging Interpolated FIR Filters [DSP Tips Tricks]." In: *IEEE Signal Processing Magazine* 24.5 (2007), pp. 140–143.

[60] J. E. Cadena and A. A. L. Beex. “Interpolated FIR based practically perfect reconstruction filter bank.” In: *2016 50th Asilomar Conference on Signals, Systems and Computers*. 2016, pp. 869–876.

[61] R. Lyons. “Interpolated narrowband lowpass FIR filters.” In: *IEEE Signal Processing Magazine* 20.1 (2003), pp. 50–57.

[62] J. L. H. Webb and D. C. Munson. “A new approach to designing computationally efficient interpolated FIR filters.” In: *IEEE Transactions on Signal Processing* 44.8 (1996), pp. 1923–1931.

[63] T. Saramaki, T. Neuvo, and S. K. Mitra. “Design of computationally efficient interpolated FIR filters.” In: *IEEE Transactions on Circuits and Systems* 35.1 (1988), pp. 70–88.

[64] G. Molnar, A. Dudarin, and M. Vucic. “Design and Multiplierless Realization of Maximally Flat Sharpened-CIC Compensators.” In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 65.1 (2018), pp. 51–55.

[65] G. Jovanovic-Dolecek and S. K. Mitral. “Efficient sharpening of CIC decimation filter.” In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)*. Vol. 6. 2003, pp. VI–385.

[66] G. Jovanovic-Dolecek and S. K. Mitra. “A new two-stage sharpened comb decimator.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 52.7 (2005), pp. 1414–1420.

[67] G. Molnar, A. Dudarin, and M. Vucic. “Minimax design of multiplierless sharpened CIC filters based on interval analysis.” In: *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2016, pp. 94–98.

[68] M. G. C. Jimenez, U. Meyer-Baese, and G. J. Dolecek. “Computationally efficient CIC-based filter with embedded Chebyshev sharpening for the improvement of aliasing rejection.” In: *Electronics Letters* 53.4 (2017), pp. 281–283.

[69] M. Laddomada, D. E. Troncoso, and G. J. Dolecek. “Improved sharpening of comb-based decimation filters: Analysis and design.” In: *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*. 2014, pp. 11–16.

[70] T. D. Memon, P. Beckett, and Z. M. Hussain. “Analysis and design of a ternary FIR filter using sigma delta modulation.” In: *2009 IEEE 13th International Multitopic Conference*. 2009, pp. 1–5.

[71] T. C. Pham et al. “Implementation of a short word length ternary FIR filter in both FPGA and ASIC.” In: *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications Computing (SigTelCom)*. 2018, pp. 45–50.

[72] S. P. Ghanekar, S. Tantaratana, and L. E. Franks. “Multiplier-free FIR filters with periodically time-varying ternary coefficients.” In: *[1991] Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems Computers*. 1991, 1037–1041 vol.2.

[73] R. Hezar and V. K. Madisetti. “Low-power digital filter implementations using ternary coefficients.” In: *VLSI Signal Processing, IX*. 1996, pp. 179–188.

[74] Woo Jin Oh and Yong Hoon Lee. “Implementation of programmable multiplierless FIR filters with powers-of-two coefficients.” In: *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* 42.8 (1995), pp. 553–556.

[75] T. D. Memon, P. Beckett, and A. Z. Sadik. “Single-Bit and Conventional FIR Filter Comparision in State-of-Art FPGA.” In: *2009 Fifth International Conference on MEMS NANO, and Smart Systems*. 2009, pp. 72–76.

[76] S. S. Yedlapalli and K. V. S. Hari. “The canonic linear-phase FIR lattice structures.” In: *2010 National Conference On Communications (NCC)*. 2010, pp. 1–5.

[77] Y. Tsao and K. Choi. “Area-Efficient VLSI Implementation for Parallel Linear-Phase FIR Digital Filters of Odd Length Based on Fast FIR Algorithm.” In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 59.6 (2012), pp. 371–375.

[78] J. Tian, G. Li, and Q. Li. “Hardware-efficient parallel structures for linear-phase FIR digital filter.” In: *2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS)*. 2013, pp. 995–998.

[79] Tian-Bo Deng. “Symmetry-based low-complexity variable fractional-delay FIR filters.” In: *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004.* Vol. 1. 2004, 194–199 vol.1.

[80] A. Kumar, S. Yadav, and N. Purohit. “Exploiting Coefficient Symmetry in Conventional Polyphase FIR Filters.” In: *IEEE Access* 7 (2019), pp. 162883–162897.

[81] S. Akhter, S. Kumar, and D. Bareja. “Design and Analysis of Distributed Arithmetic based FIR Filter.” In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. 2018, pp. 721–726.

[82] S. Khan and Z. A. Jaffery. “Low power FIR filter implementation on FPGA using parallel Distributed Arithmetic.” In: *2015 Annual IEEE India Conference (INDICON)*. 2015, pp. 1–5.

[83] N. J. Grande and S. Sridevi. “ASIC implementation of shared LUT based distributed arithmetic in FIR Filter.” In: *2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS)*. 2017, pp. 1–4.

[84] S. Ruth Joanna and A. V. Anathalakshmi. “Design and implementation of efficient adaptive FIR filter based on distributed arithmetic.” In: *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. 2015, pp. 1–4.

[85] P. L. J. Raj and T. Vigneswaran. “A paradigm of distributed arithmetic (DA) approaches for digital FIR filter.” In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. 2016, pp. 4668–4672.

[86] A. Kurosu et al. “A technique to truncate IIR filter impulse response and its application to real-time implementation of linear-phase IIR filters.” In: *IEEE Transactions on Signal Processing* 51.5 (2003), pp. 1284–1292.

[87] A. Wang and J. O. Smith. “Some properties of tail-canceling IIR filters.” In: *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*. 1997, 4 pp.

[88] A. Wang and J. O. Smith. “On fast FIR filters implemented as tail-canceling IIR filters.” In: *IEEE Transactions on Signal Processing* 45.6 (1997), pp. 1415–1427.

[89] L. Horowitz. “The effects of spline interpolation on power spectral density.” In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 22.1 (1974), pp. 22–27.

[90] Peter Nuijten and Derk Reefman. “Why Direct Stream Digital (DSD) is the best choice as a digital audio format.” In: *Journal of the Audio Engineering Society* (May 2001).

[91] James A. S. Angus. “A New Method of Applying High Levels of Dither to Delta-Sigma Modulators.” In: *Audio Engineering Society Convention 117*. Oct. 2004. URL: http://www.aes.org/e-lib/browse.cfm?elib=12953.

[92] Wai Laing Lee. “A novel higher order interpolative modulator topology for high resolution oversampling A/D converters.” In: *Master’s thesis, Massachusetts Institute of Technology* (1987).

[93] Tapani Ritoniemi, Teppo Karema, and Hannu Tenhunen. “Design of stable high order 1-bit sigma-delta modulators.” In: *IEEE International Symposium on Circuits and Systems*. IEEE. 1990, pp. 3267–3270.

[94] S. Plekhanov, I. A. Shkolnikov, and Y. B. Shtessel. "High order sigma-delta modulator design via sliding mode control." In: *Proceedings of the 2003 American Control Conference, 2003.* Vol. 1. 2003, 897–902 vol.1.

[95] Ivar Lokken et al. "Quantizer nonoverload criteria in sigma–delta modulators." In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 53.12 (2006), pp. 1383–1387.

[96] A. Sanyal and N. Sun. "An enhanced ISI shaping technique for multi-bit $\Delta\Sigma$ DACs." In: *2014 IEEE International Symposium on Circuits and Systems (ISCAS).* June 2014, pp. 2341–2344. DOI: 10.1109/ISCAS.2014.6865641.

[97] Vincent O'Brien. "Design of high order mismatch and ISI shaping dynamic element matching decoders for delta sigma data converters." In: 2017.

[98] Derk Reefman et al. "A New Digital-to-Analogue Converter Design Technique for HiFi Applications." In: *Audio Engineering Society Convention 114.* Mar. 2003. URL: http://www.aes.org/e-lib/browse.cfm?elib=12538.

[99] Chuan-Hung HSIAO, Sung-Han WEN, and Kuan-Ta CHEN. *HIGH LINEARITY DIGITAL-TO-ANALOG CONVERTER WITH ISI-SUPPRESSING METHOD.* Jan. 2020.

[100] J. Remple and I. Galton. "The Effects of Inter-Symbol Interference in Dynamic Element Matching DACs." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.1 (2017), pp. 14–23.

[101] Aria Eshraghi, Ramkishore Ganti, and Weinan Gao. *High performance delta sigma ADC using a feedback NRZ sin DAC.* US Patent 6,462,687. Oct. 2002.

[102] K. Lee, M. Bonu, and G. C. Temes. "Noise-coupled /spl Delta//spl Sigma/ ADC's." In: *Electronics Letters* 42.24 (2006), pp. 1381–1382.

[103] T. Ritoniemi, T. Karema, and H. Tenhunen. "Design of stable high order 1-bit sigma-delta modulators." In: *IEEE International Symposium on Circuits and Systems.* 1990, 3267–3270 vol.4.

[104] Jason Remple and Ian Galton. "The effects of inter-symbol interference in dynamic element matching DACs." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.1 (2016), pp. 14–23.

[105] Arindam Sanyal, Long Chen, and Nan Sun. "Dynamic Element Matching With Signal-Independent Element Transition Rates for Multibit $\Delta\Sigma$ Modulators." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 62.5 (2015), pp. 1325–1334.

[106] Ramy Saad, Sebastian Hoyos, and Samuel Palermo. “Analysis and modeling of clock-jitter effects in Delta-Sigma modulators.” In: *MATLAB—A Fundamental Tool for Scientific Computing and Engineering Applications*. Vol. 1. InTech, 2012, pp. 393–422.

[107] J.P. Colinge. *FinFETs and Other Multi-Gate Transistors*. Integrated Circuits and Systems. Springer, 2008. ISBN: 9780387717517. URL: https://books.google.it/books?id=t1ojkCdTGEEC.

[108] T. C. Leslie and B. Singh. “An improved sigma-delta modulator architecture.” In: *IEEE International Symposium on Circuits and Systems*. 1990, 372–375 vol.1.

[109] Y. Tang, X. Chen, and H. Zhu. “A 108-dB SNDR 2–1 MASH ΔΣ Modulator with First-Stage Multibit for Audio Application.” In: *2018 IEEE 3rd International Conference on Integrated Circuits and Microsystems (ICICM)*. 2018, pp. 336–340.

[110] N. Maghari and Un-Ku Moon. “Multi-loop efficient sturdy MASH delta-sigma modulators.” In: *2008 IEEE International Symposium on Circuits and Systems*. 2008, pp. 1216–1219.

[111] Changsok Han, Taewook Kim, and N. Maghari. “SMASH-MASH delta-sigma modulator using noise-shaping quantizers.” In: *2016 14th IEEE International New Circuits and Systems Conference (NEWCAS)*. 2016, pp. 1–4.

[112] K. Seo et al. “An incremental zoom sturdy MASH ADC.” In: *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. 2017, pp. 1013–1016.

[113] W. Jin and K. Pun. “A DEM-Free Sturdy MASH Delta-Sigma Modulator with a Highly-Linear Tri-level DAC.” In: *2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*. 2019, pp. 1–2.

[114] N. Maghari et al. “Mixed-Order Sturdy MASH Δ-Σ Modulator.” In: *2007 IEEE International Symposium on Circuits and Systems*. 2007, pp. 257–260.

[115] S. Pamarti and I. Galton. “LSB Dithering in MASH Delta–Sigma D/A Converters.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 54.4 (2007), pp. 779–790.

[116] Victor Gonzalez-Diaz et al. “Efficient Dithering in MASH Sigma-Delta Modulators for Fractional Frequency Synthesizers.” In: *IEEE Trans. on Circuits and Systems* 57-I (Jan. 2010), pp. 2394–2403.

[117] R. Hezar et al. “A 110dB SNR and 0.5mW current-steering audio DAC implemented in 45nm CMOS.” In: *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*. 2010, pp. 304–305.

[118] P. S. Crovetti. “All-Digital High Resolution D/A Conversion by Dyadic Digital Pulse Modulation.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.3 (2017), pp. 573–584.

[119] M. Usmonov et al. “Suppression of Quantization-Induced Limit Cycles in Digitally Controlled DC-DC Converters by Dyadic Digital Pulse Width Modulation.” In: *2019 IEEE Energy Conversion Congress and Exposition (ECCE)*. 2019, pp. 2224–2231.

[120] R. Cellier et al. “An review of fully digital audio class D amplifiers topologies.” In: *2009 Joint IEEE North-East Workshop on Circuits and Systems and TAISA Conference*. 2009, pp. 1–4.

[121] S. Luo and D. Li. “A Sixth-Order PWM Modulator for Digital Input Class-D Audio Amplifiers.” In: *2013 International Conference on Computational and Information Sciences*. 2013, pp. 1253–1256.

[122] Rolf Esslinger, Gerhard Gruhler, and R. W. Stewart. “Feedback Strategies in Digitally Controlled Class-D Amplifiers.” In: *Audio Engineering Society Convention 114*. Mar. 2003. URL: http://www.aes.org/e-lib/browse.cfm?elib=12570.

[123] S. O. Aase. “Digital removal of pulse-width-modulation-induced distortion in class-D audio amplifiers.” In: *IET Signal Processing* 8.6 (2014), pp. 680–692.

[124] F. Chierchie and S. O. Aase. “Volterra Models for Digital PWM and Their Inverses.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 62.10 (2015), pp. 2606–2616.

[125] Lars Risbo and Thomas Mørch. “Performance of an All-Digital Power Amplification System.” In: *Audio Engineering Society Convention 104*. May 1998. URL: http://www.aes.org/e-lib/browse.cfm?elib=8485.

[126] F. Chierchie and E. E. Paolini. “Digital Distortion-Free PWM and Click Modulation.” In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 65.3 (2018), pp. 396–400.

[127] L. Stefanazzi, A. R. Oliva, and E. E. Paolini. “Alias-Free Digital Click Modulator.” In: *IEEE Transactions on Industrial Informatics* 9.2 (2013), pp. 1074–1083.

[128] T. Domingues, M. Santos, and G. Tavares. “A Click Modulation audio player.” In: *2015 Conference on Design of Circuits and Integrated Systems (DCIS)*. 2015, pp. 1–6.

[129] L. Stefanazzi et al. “Low Distortion Switching Amplifier With Discrete-Time Click Modulation.” In: *IEEE Transactions on Industrial Electronics* 61.7 (2014), pp. 3511–3518.

[130] P. Wagh. “Closed-form spectral analysis of pulse-width modulation.” In: *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems (Cat. No.01CH37196)*. Vol. 3. 2001, 799–802 vol. 2.

[131] T. Domingues, M. Santos, and G. Tavares. “Low frequency PWM modulation for high efficiency Class-D audio driving.” In: *Design of Circuits and Integrated Systems*. 2014, pp. 1–5.

[132] A. J. Magrath and M. B. Sandler. “Hybrid pulse width modulation/sigma-delta modulation power digital-to-analogue converter.” In: *IEE Proceedings - Circuits, Devices and Systems* 143.3 (1996), pp. 149–156.

[133] E. Hogenauer. “An economical class of digital filters for decimation and interpolation.” In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2 (1981), pp. 155–162.

[134] M. Hyder and K. Mahata. “An approximate L0 norm minimization algorithm for compressed sensing.” In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 3365–3368.

[135] T. T. Nguyen et al. “NP-hardness of ℓ0 minimization problems: revision and extension to the non-negative setting.” In: *2019 13th International conference on Sampling Theory and Applications (SampTA)*. 2019, pp. 1–4.

[136] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. “Atomic decomposition by basis pursuit.” In: *SIAM Journal on Scientific Computing* 20 (1998), pp. 33–61.

[137] Anita Thengade and Rucha Dondal. “Genetic Algorithm – Survey Paper.” In: *IJCA Proc National Conference on Recent Trends in Computing, NCRTC* 5 (Jan. 2012).

[138] J. Kennedy and R. Eberhart. “Particle swarm optimization.” In: *Proceedings of ICNN’95 - International Conference on Neural Networks*. Vol. 4. 1995, 1942–1948 vol.4.

[139] P.J. van Laarhoven and E.H. Aarts. *Simulated Annealing: Theory and Applications*. Mathematics and Its Applications. Springer Netherlands, 1987. ISBN: 9789027725134. URL: https://books.google.it/books?id=-IgUab6Dp%5C_IC.

[140] Zong Woo Geem, Joong Hoon Kim, and Gobichettipalayam Vasudevan Loganathan. “A new heuristic optimization algorithm: harmony search.” In: *simulation* 76.2 (2001), pp. 60–68.

[141] Dervis Karaboga and Bahriye Basturk. “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm.” In: *Journal of global optimization* 39.3 (2007), pp. 459–471.

[142] Philip E Gill and Walter Murray. “Algorithms for the solution of the nonlinear least-squares problem.” In: *SIAM Journal on Numerical Analysis* 15.5 (1978), pp. 977–992.

[143] S. Nishimura, K. Hirano, and R. Pal. “A new class of very low sensitivity and low roundoff noise recursive digital filter structures.” In: *IEEE Transactions on Circuits and Systems* 28.12 (1981), pp. 1152–1158.

[144] Wilhelm Werner. “Polynomial interpolation: Lagrange versus newton.” In: *Mathematics of computation* (1984), pp. 205–217.

[145] Alok Dutt, Ming Gu, and Vladimir Rokhlin. “Fast algorithms for polynomial interpolation, integration, and differentiation.” In: *SIAM Journal on Numerical Analysis* 33.5 (1996), pp. 1689–1711.

[146] Yu N Subbotin. “Piecewise-polynomial (spline) interpolation.” In: *Mathematical notes of the Academy of Sciences of the USSR* 1.1 (1967), pp. 41–45.

[147] Michael Revers. “A Survey on Lagrange Interpolation Based on Equally Spaced Nodes.” In: *Advanced Problems in Constructive Approximation.* Ed. by Martin D. Buhmann and Detlef H. Mache. Basel: Birkhäuser Basel, 2003, pp. 153–164. ISBN: 978-3-0348-7600-1.

[148] S. A. Martucci. “Symmetric convolution and the discrete sine and cosine transforms.” In: *IEEE Transactions on Signal Processing* 42.5 (1994), pp. 1038–1051.

[149] Huibert Kwakernaak and Raphael Sivan. “Modern signals and systems.” In: *STIA* 91 (1991), p. 11586.

[150] M Richardson. “Fundamentals of the discrete Fourier transform.” In: *Sound and vibration magazine* (1978), pp. 40–46.

[151] Anil K Jain. “A sinusoidal family of unitary transforms.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4 (1979), pp. 356–365.

[152] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. “Discrete cosine transform.” In: *IEEE transactions on Computers* 100.1 (1974), pp. 90–93.

[153] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications.* Academic press, 2014.

[154] Andrew B Watson. "Image compression using the discrete cosine transform." In: *Mathematica journal* 4.1 (1994), p. 81.

[155] AM Raid et al. "Jpeg image compression using discrete cosine transform-A survey." In: *arXiv preprint arXiv:1405.6147* (2014).

[156] Alexandre Balkanski et al. *System for compression and decompression of video data using discrete cosine transform and coding techniques.* US Patent 5,253,078. Oct. 1993.

[157] Markus Puschel and José MF Moura. "Algebraic signal processing theory: Cooley–Tukey type algorithms for DCTs and DSTs." In: *IEEE Transactions on Signal Processing* 56.4 (2008), pp. 1502–1521.

[158] Frederick N Fritsch and Ralph E Carlson. "Monotone piecewise cubic interpolation." In: *SIAM Journal on Numerical Analysis* 17.2 (1980), pp. 238–246.

[159] Hiroshi Akima. "A new method of interpolation and smooth curve fitting based on local procedures." In: *Journal of the ACM (JACM)* 17.4 (1970), pp. 589–602.

[160] Carl De Boor et al. *A practical guide to splines.* Vol. 27. springer-verlag New York, 1978.

[161] A. W. Paeth. "A Fast Algorithm for General Raster Rotation." In: *Proceedings of Graphics Interface and Vision Interface '86.* GI '86. Vancouver, British Columbia, Canada: Canadian Man-Computer Communications Society, 1986, pp. 77–81. URL: http://graphicsinterface.org/wp-content/uploads/gi1986-15.pdf.

[162] Manjeet Kumar and Tarun Kumar Rawat. "Design of fractional order differentiator using type-III and type-IV discrete cosine transform." In: *Engineering Science and Technology, an International Journal* 20.1 (2017), pp. 51–58. ISSN: 2215-0986. DOI: https://doi.org/10.1016/j.jestch.2016.07.002. URL: http://www.sciencedirect.com/science/article/pii/S2215098615300197.

[163] Mohammed-Salah Abdelouahab and Nasr-Eddine Hamri. "The Grünwald-Letnikov fractional-order derivative with fixed memory length." In: *Mediterranean Journal of Mathematics* 13.2 (2016), pp. 557–572.

[164] Rudy J Van De Plassche. "Dynamic element matching for high-accuracy monolithic D/A converters." In: *IEEE Journal of solid-state Circuits* 11.6 (1976), pp. 795–800.

[165] Ian Galton. "Why dynamic-element-matching DACs work." In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 57.2 (2010), pp. 69–74.

[166] RK Henderson and OJAP Nys. “Dynamic element matching techniques with arbitrary noise shaping function.” In: *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96.* Vol. 1. IEEE. 1996, pp. 293–296.

[167] Bruce Duewer, Heling Yi, and John Melanson. “A Multi-bit Delta-Sigma DAC with Mismatch Shaping in the Feedback Loop.” In: *Audio Engineering Society Convention 115.* Oct. 2003. URL: http://www.aes.org/e-lib/browse.cfm?elib=12392.

[168] Akira Yasuda, Hiroshi Tanimoto, and Tetsuya Iida. “A third-order/spl Delta/-/spl Sigma/modulator using second-order noise-shaping dynamic element matching.” In: *IEEE journal of solid-state circuits* 33.12 (1998), pp. 1879–1886.

[169] Derk Reefman and Erwin Janssen. “Enhanced sigma delta structures for super audio CD applications.” In: *Audio Engineering Society Convention 112.* Audio Engineering Society. 2002.

[170] C. Basetas, N. Temenos, and P. P. Sotiriadis. “Implementation of Multi-Step Look-Ahead Sigma-Delta Modulators Using IC Technology.” In: *2018 IEEE International Frequency Control Symposium (IFCS).* 2018, pp. 1–5.

[171] C. Basetas, T. Orfanos, and P. P. Sotiriadis. “A Class of 1-Bit Multi-Step Look-Ahead Σ - Δ Modulators.” In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.1 (2017), pp. 24–37.

[172] T. Hayashi et al. “A multistage delta-sigma modulator without double integration loop.” In: *1986 IEEE International Solid-State Circuits Conference. Digest of Technical Papers.* Vol. XXIX. 1986, pp. 182–183.

[173] N. Maghari et al. “Sturdy MASH /spl Delta//spl Sigma/ modulator.” In: *Electronics Letters* 42.22 (2006), pp. 1269–1270.

[174] David K Su and Bruce A Wooley. “A CMOS oversampling D/A converter with a current-mode semidigital reconstruction filter.” In: *IEEE journal of solid-state circuits* 28.12 (1993), pp. 1224–1233.

[175] H. Mouton and B. Putzeys. “Understanding the PWM Nonlinearity: Single-Sided Modulation.” In: *IEEE Transactions on Power Electronics* 27.4 (2012), pp. 2116–2128.

[176] *Terasic DE1-SoC Board.* URL: https://www.terasic.com.tw/cgi-bin/page/archive.pl?Language=English&CategoryNo=165&No=836.

[177] Chiara Bartolozzi et al. “Live demonstration: Tactile events from off-the-shelf sensors in a robotic skin.” In: *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS).* IEEE. 2017, pp. 1–1.

[178] Alberto Paltrinieri et al. "Approximate-Computing Architectures for Motion Estimation in HEVC." In: *2018 New Generation of CAS (NGCAS)*. IEEE. 2018, pp. 190–193.

[179] Alberto Paltrinieri et al. "On the effect of approximate-computing in motion estimation." In: *Journal of Low Power Electronics* 15.1 (2019), pp. 40–50.

[180] Paolo Selvo et al. "An optimized partial-distortion-elimination based sum-of-absolute-differences architecture for high-efficiency-video-coding." In: *International Conference on Applications in Electronics Pervading Industry, Environment and Society*. Springer. 2018, pp. 245–251.

[181] Paolo Selvo et al. "An Optimized Partial-Distortion-Elimination Based Sum-of-Absolute-Differences." In: *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2018* 573 (2019), p. 245.

[182] Maurizio Capra et al. "Edge computing: A survey on the hardware requirements in the internet of things world." In: *Future Internet* 11.4 (2019), p. 100.

[183] Erik Anzalone et al. "Low-Power Hardware Accelerator for Sparse Matrix Convolution in Deep Neural Network." In: *Progresses in Artificial Intelligence and Neural Systems*. Springer, 2020, pp. 79–89.

[184] Luigi Sole et al. "VLSI Architectures for the Steerable-Discrete-Cosine-Transform (SDCT)." In: *International Conference on Applications in Electronics Pervading Industry, Environment and Society*. Springer. 2019, pp. 137–143.

[185] Riccardo Peloso et al. "Steerable-Discrete-Cosine-Transform (SDCT): Hardware Implementation and Performance Analysis." In: *Sensors* 20.5 (2020), p. 1405.