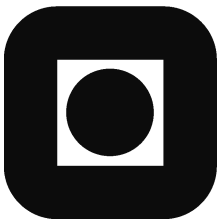
Digital-to-Analog Conversion in High Resolution Audio

by

Ivar Løkken

September 2008

N T N U



Submitted to the Norwegian University of Science and Technology
in partial fulfilment of the requirements for the degree philosophiae doctor

NTNU
Norwegian University of Science and Technology

Thesis for the degree philosophiae doctor
Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Electronics and Telecommunications

© Ivar Løkken

ISBN 978-82-471-1209-0 [printed version]
ISBN 978-82-471-1210-6 [electronic version]
ISSN 1503-8181

NTNU Doctoral Theses 2008:257

Printed by NTNU-trykk

Abstract

This thesis describes theoretical and simulation-based work on digital-to-analog conversion for high resolution audio. The emphasis of the work has been exploration and clarification of issues of contention in previous art. The work has resulted in five scientific papers published in international peer-reviewed journals and conference proceedings, and these papers constitute the main contribution. The papers are included as appendixes, whereas the preceding monograph serves to provide the necessary background for understanding the results, and also their relevance in an audio context. It should be noted that although the research primarily treats DA conversion, the findings and conclusions are largely transferable also to AD conversion since audio ADC performance is often limited by its usually compulsory feedback DAC.

The first paper, published in the *Journal of the Audio Engineering Society*, explores power modulation of the quantization error and the need for dithering in delta-sigma modulators. There has been a lot of dispute on this issue; previous publications having both argued that the DSM is self-dithering and that it has the same dither requirements as a regular REQ. By exploring noise power modulation in the baseband it is shown that even high order DSMs are not self-dithering in the true sense, but that the adverse effects of quantization are reduced when the loop filter is of high order. If the REQ is multi-bit the noise power modulation can be made negligible compared to any practical levels of circuit noise.

The second paper, published in *IEEE Transactions on Circuits and Systems Part II*, explores a class of DSM called non-overloading or NOL modulators. Designing the DSM to be NOL is the only known way to *guarantee* stability for high order loops, and also the only way to guarantee no quantization noise power modulation. The paper proves that NOL design criteria are equivalent for OF and EF modulators, repudiating a claim of difference in a previous publication, and also their equivalence for rounding and truncating quantizers. Although the results are developed for a certain class of modulators, the methods are easily generalized to any DSM design. It is found desirable to use a many-bit REQ since a NOL DSM with good input swing is then allowed.

The third paper, presented at the *31st Conference of the Audio Engineering Society*, shows a useful utilization of the results developed in the second paper. Using a many-bit DSM is desirable for several reasons, but will in straightforward implementation require a DEM network of excessive complexity. A previously proposed method to circumvent this is to segment the DAC and DEM using a dedicated Segmentation-DSM. Previous art has used SDSMs with a FIR loop filter to ensure no DAC saturation, restricting the concept to very non-optimal designs. This publication utilizes the NOL method to design IIR SDSMs with significantly improved performance.

The fourth paper, submitted to *Analog Integrated Circuits and Signal Processing*, describes the development of simplified estimates for DSM DAC errors. The mathematical treatment of high order DSMs is exceedingly difficult, but simplifications and rules of thumb have been developed that enable design engineers to make quite straightforward optimization of relevant DSM parameters. A major drawback is that these approximations do not account for analog error sources in the DAC and may therefore lead to unfortunate design choices. This paper explores how common DAC errors depend on the DSM transfer function, and presents extensions of known approximation methods to also include the impact the DSM has on DAC waveform distortion. Again it is confirmed that using a many-bit DSM is advantageous, and also that a conservative DSM design will make the DAC less susceptible to errors.

The fifth and last paper, presented at the *124th Convention of the Audio Engineering Society*, utilizes the methods presented in the fourth paper to optimize a DAC with regards to jitter noise. Clock jitter is one of the most critical performance bottlenecks in high resolution audio, and the paper proposes ways to minimize the DAC's jitter susceptibility. The simplified approximation methods are employed and extended to show that a semidigital FIR DAC gives a more benign output waveform than a segmented DEM DAC of comparable complexity, and that it will be a preferable solution if jitter dominates the error budget. A simple method is also shown to estimate effects of implementation inaccuracies in the analog filter coefficients.

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree Philosophiae Doctor (Ph.D.) at the Norwegian University of Science and Technology (NTNU), Department of Electronics and Telecommunications. The work has been funded by the Norwegian Research Council under grant 162101 SPECK. My main supervisor has been Professor Trond Sæther and my co-supervisor has been Dr.Ing. Bjørnar Hernes.

The studies have been carried out in the period from January 2005 to May 2008. The work includes the equivalent of a little over half a year of full-time course studies (39 ECTS credits), teaching one fifth year master level course on data converters and supervising two students on their master projects. The first year and a half was spent at the university, and after this the work was carried out at the Nordic Semiconductor data converter department. When the Nordic data converter business was acquired by Chipidea Microelectronica, we were two Ph.D. students enrolled at the department who joined in on the move. Chipidea Norway sadly turned out to be a short lived venture, as irreconcilably different visions for the future between the mother company and the Norwegian department staff, eventually led to all regular employees in Norway leaving in order to pursue a fresh start with new company Arctic Silicon Devices. I and the other Ph.D. student were still allowed to finish our studies at the university where I am now, while remaining on Chipidea contracts. So in many ways the circle seems complete, as I am now finishing this study more than three years later, back where it started.

Acknowledgements

This thesis is dedicated to my friends and family, in particular my parents and my sister. If it wasn't for you, I would not have been here today.

As I am sitting in my office at the university writing this, spring is in the air and almost four years have passed since I was in a similar situation as a master student. Then too Trond was my supervisor and in our meetings, which often slanted into hi-fi talks since we are both devout audiophiles and music fans, he aired the idea of doing a Ph.D. on audio conversion. With my background and interests I found this to be an extremely exciting opportunity and jumped at it, and it was arranged between the university, Nordic and the research council to initiate a project. Four years later I have no regrets, as I have learned a lot and met many interesting people on the way. I would like to express my gratitude to those who have been of help; both professionally and administratively, during the years spent pursuing this degree.

First and foremost my colleague and fellow Ph.D. student Anders Vinje, who has travelled the same slightly bumpy path to this point as me, has been of tremendous aid during the work on my papers and as a general "sparring partner" during problem solving. His mathematical prowess easily surpasses mine and I would have had a lot more troubled times in front of the notebook if it wasn't for him. Hurdles related to our somewhat unusual work situation have also been easier to climb than if I had been in it alone.

I am also in debt to my main supervisor Trond and co-supervisor Bjørnar, who have been of great help administratively. The path on which I set out is one they were both familiar with from when they did their own Ph.D. degrees, and they helped me to keep my focus on the light in the end of the tunnel even when at times the tunnel itself seemed dark.

To spend one and a half years with the former Nordic data converter team gave a sobering insight into the practical aspects of integrated circuit design. Listening in on their project meetings provided a pragmatic context to my theoretical work and helped me maintain a meaningful focus. Thanks to Morten Dammen, Terje Granum, Øystein Moldsvor, Christian Holdø, Håvard Korsvoll, Frode Telstø, Terje Andersen, and Atle Briskemyr for providing an informal and positive work environment with a lot of expertise. Johnny Bjørnsen who remained at Nordic after the takeover also deserves mention as part of this group.

Managing a master level course on data converters and advising two master students on their projects was invaluable in learning to communicate and share my knowledge. I thank Lasse Olsen and Florian Bousquet for being interested and skilful students.

I also want to thank the staff and students at the NTNU Circuits and Systems group for my time and discussions with them, in particular Prof. Trond Ytterdal, Carsten Wulff, and Rune Kaald. And last but not least; thank you to the management at Chipidea and Prof. João Vital for allowing me and Anders to finish our degrees even though the data converter team fell apart. I am sorry the endeavour of Chipidea Norway ended the way it did.

Finally, I would like to pay a small homage to the late audio-guru Steen Aage Duelund, with whom I got to discuss and learn from just before he untimely went away. Although his fields of expertise were only tangentially related to mine, his clairvoyance and ability to think outside the box, his holistic philosophical approach to the art of audio, and his great enthusiasm provided me with a lot of inspiration. His presence is missed.

Trondheim, May 2008

Ivar Løkken.

Table of Contents

List of Figures

List of Abbreviations

AAF	Anti-Alias Filter
ABE	Analog Back End
ADC	Analog to Digital Converter (or: Analog to Digital Conversion)
ADDA	Analog-Digital-Digital-Analog
AES	Audio Engineering Society
AFE	Analog Front End
ANSI	American National Standards Institute
ARA	Acoustic Renaissance for Audio
ASRC	Asynchronous Sample-Rate Converter
BIBO	Bounded Input Bounded Output
BJT	Bipolar Junction Transistor
CD	Compact Disc
CF	Characteristic Function
CMOS	Complementary Metal Oxide Semiconductor
DAC	Digital to Analog Converter (or: Digital to Analog Conversion)
DB:	DeciBel
DBFS	DeciBel relative to full-scale
DCT	Direct Charge Transfer
DEM	Dynamic Element Matching
DFT	Discrete Fourier Transform
DIN	Deutsches Institut für Normung
DNL	Differential Non-Linearity
DSM	Delta Sigma Modulator (or: Delta Sigma Modulation)
DSD	Direct Stream Digital
DSP	Digital Signal Processing
DTFT	Discrete Time Fourier Transform
DVD	Digital Versatile Disc
DVD-A	DVD-Audio
DWA	Data Weighted Averaging
EF	Error Feedback
ENOB	Effective Number of Bits
FET	Field Effect Transistor
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FOM	Figure of Merit
FPGA	Field Programmable Gate Array
FS	Full-Scale (or: If written f_s; sampling frequency)
HD	Harmonic Distortion
HD2	Second Harmonic Distortion
HD3	Third Harmonic Distortion
HF	High Frequency
Hi-res	High Resolution
IC	Integrated Circuit
IEEE	Institute of Electrical and Electronics Engineers
IEC	International Electrotechnical Commission
IFIR	Interpolated FIR
IIR	Infinite Impulse Response

ILA	Individual Level Averaging
INL	Integral Non-Linearity
ISI	Inter Symbol Interference
ISO	International Standardizing Organization
JTF	Jitter Transfer Function
LF	Low Frequency
LFSR	Linear Feedback Shift Register
LPCM	Linear Pulse Code Modulation
LSB	Least Significant Bit
MAC	Multiplier Accumulator
MASH	Multi stAge noise-SHaping
MOS	Metal Oxide Semiconductor
MSB	Most Significant Bit
MSE	Mean Square Error
NOL	Non-Overloading
NOS	Non-Oversampling
NRZ	Non Return to Zero
NTF	Noise Transfer Function
OF	Output Feedback
OSR	Oversampling Ratio
PCM	Pulse Code Modulation
PDF	Probability Density Function
PDM	Pulse Density Modulation
PLL	Phase Locked Loop
PRNG	Pseudo Random Number Generator
PSD	Power Spectral Density
PWM	Pulse Width Modulation
RCF	Reconstruction Filter
REQ	Re-Quantizer (or: Re-Quantization)
RMS	Root Mean Square
ROC	Region of Convergence
RZ	Return to Zero
SACD	Super Audio Compact Disc
SC	Switched Capacitor
SFDR	Spurious Free Dynamic Range
SJNR	Signal to Jitter Noise Ratio
SMNR	Signal to Mismatch Noise Ratio
SNDR	Signal to Noise and Distortion Ratio
SNR	Signal to Noise Ratio
SP-DIF	Sony/Philips Digital Interface Format
SPL	Sound Pressure Level
SQNR	Signal to Quantization Noise Ratio
SSNR	Signal to Switching Noise Ratio
STF	Signal Transfer Function
THD	Total Harmonic Distortion
THD+N	Total Harmonic Distortion and Noise
TNSM	Trellis Noise-Shaping Modulator
VLSI	Very Large Scale Integration
VQ	Vector Quantization

Chapter 1

Introduction

1.1 Hearing and Audio Quality

When Edison invented the phonograph in the 1870s [1], he probably didn't envision what a major industry the recording, conservation and reproduction of music would become. Advances in technology have steadily increased the performance as well as availability of reproduced sound, and a listener can now fit an entire music library in transparent quality into his pocket.

In development of audio technology, the qualitative context is represented by understanding and knowledge of the human auditory system and its properties. Fletcher and Munson did important early work in quantifying the bandwidth and sensitivity of the human hearing [2], which resulted in the equal loudness contour and the *phon* denomination of perceived loudness. The Fletcher-Munson curves were later revised as the Robinson-Dadson curves [3], which became the basis of the ISO226 equal loudness standard.

Figure 1 shows the equal loudness curves according to ISO226. The 0-phon curve is known as the *threshold of audibility* and the 120-phon curve as the *threshold of pain*. The span between these two thresholds is generally acknowledged as the usable *dynamic range* of the human auditory system. It thus represents a measure for the desirable dynamic range in audio equipment. The y-axis is the absolute SPL in dB relative to a reference of 20µPa RMS.

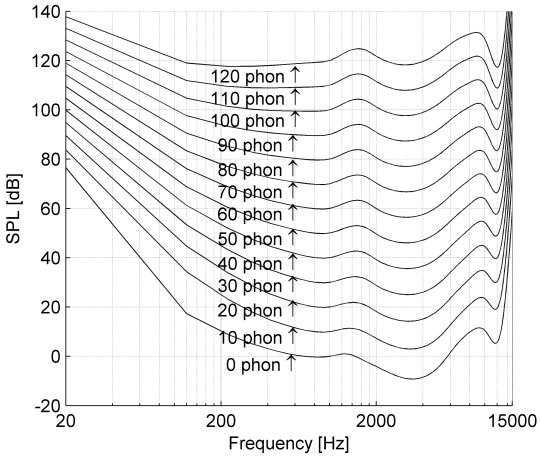


Figure 1: Equal loudness contours (ISO226)

In the frequency range of approximately 2kHz to 5kHz, called the midrange, the dynamic range exceeds 120dB. It is well maintained into the lower (bass) and higher (treble) frequency regions, but below 100 Hz and above 10 kHz it reduces significantly. The bandwidth of the hearing will vary from person to person, but the normal convention is to assume 20Hz to 20kHz for young, healthy people. Studies exist though suggesting that the way we perceive the timbre of a sound is affected by significantly higher frequencies than this [4]-[5]. Many musical instruments also have larger bandwidth than 20kHz [6].

Because of the large variation with frequency in our hearing sensitivity, uniform frequency weighting can give misleading figures when measuring audio quality. A widely accepted frequency weighting norm for sound measurement is the so-called A-weighting function (IEC/CD1672), which approximates the inverse of the 40-phon curve using six poles and four differentiating zeros. The frequency response of the standardized A-weighting function is shown in fig.2.

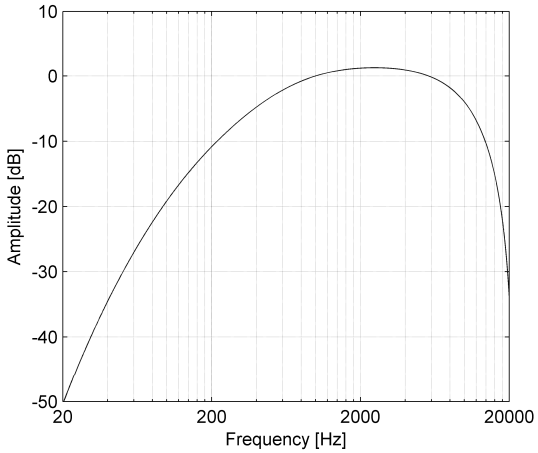


Figure 2: A-weighting function (IEC/CD 1672)

A-weighting is frequently used in specification and measurement of audio equipment including audio data converters. For instance noise is often A-weighted when measuring SNR. A predominantly white noise spectrum is reduced in power by around 3dB in the range 20Hz to 20kHz from A-weighting, meaning A-weighted SNR values are approximately 3dB better than unweighted ones, given white or predominantly white noise and this bandwidth. Based on the known characteristics of the human auditory system, the ARA commission in 1995 suggested that a high resolution audio carrier capable of full transparency should have at least 120dB dynamic range and 26kHz usable bandwidth [7]. It should be noted though that terms like "transparency" and "audio quality" are subject to an ongoing dispute between two lairs – the so-called "objectivist" and "subjectivist" factions – within the hi-fi community [8]. The "subjectivists" are generally sceptical to the authority of empirical data, and will often use arguments of solipsist and/or panpsychist nature to contend the truisms of established science. As a scientific document this thesis is founded in the "objectivist" point of view without any further discussion thereof.

1.2 A Brief Historical Review of Digital Audio

When audio entered the digital world where storage and processing capabilities increase exponentially with time as predicted by Moore [9], it rapidly became feasible to process digital audio carriers exceeding the transparency requirements defined by ARA. Practical considerations and standardization efforts have however led to a more erratic increase in de facto performance than the feasibility limits governed by Moore's Law.

Digital audio was brought to the consumer with the introduction of the Compact Disc audio playback system in the early 1980s [10]. Marketed under the pretentious slogan "*Perfect Sound Forever*", the CD-format offered 20kHz bandwidth and 96dB dynamic range in stereo. The first commercial CD-players; Sony CDP101 (Japan) and Philips CD100 (Europe), featured around 90dB dynamic range.

A complete digital audio chain will look approximately like fig.3. An instrument emits sound to an acoustoelectric transducer or microphone and the resulting electric signal is amplified and filtered by an AFE. It is then converted to digital data with an ADC, before the data is stored on a CD or other digital audio medium. During playback the medium is read and output data is transformed back to an analog signal with a DAC, amplified and filtered by an ABE and converted to sound through an electroacoustic transducer or loudspeaker. Ideally this entire process should be audibly transparent.

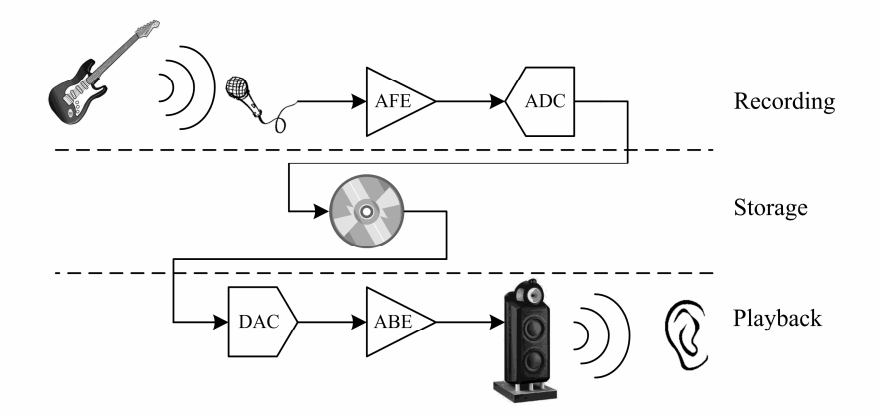


Figure 3: Digital audio recording and playback chain

It is well known that the electroacoustic transducers introduce more distortion than the other elements in the chain. Nevertheless the development process usually aspires to achieve *local transparency*, so that the component in question can be disregarded as an error source when evaluating the system. The CD-format's transparency is questionable both in terms of dynamic range and bandwidth, and its limitation to two channels makes spatial transparency unobtainable [11]. Still the CD-system has proved to be very resilient. Part of the reason for this must be attributed to the fact that it took many years before converter technology reached a level where ADC and DAC performance approached the fundamental limits of the format.

Entering the 1990s, the effective resolution of ADCs and DACs began to reach a plateau where the CD-format itself limited the performance of the ADDA process [12]. This led to an emerging demand of and research activity into higher resolution carriers, including the mentioned ARA study. By the turn of the century, two competing bids for the next generation audio carrier were launched: Philips and Sony – the companies behind the CD success – fronted the SACD [13] as its heir, whereas the working group behind the then already highly successful DVD video standard promoted the audio-specific DVD-A [14].

SACD is based on DSD; a radical 1-bit noise-shaped storage format theoretically facilitating the abolition of non-linear ADC and DAC units. It features 120dB dynamic range and 100kHz bandwidth in up to six channels. The DVD-A format uses more conventional 24-bit LPCM storage and offers a theoretical dynamic range of 144dB. The bandwidth can be up to 96kHz in two channels or 48kHz in five channels. Double-blind listening tests have failed to prove any audible differences between the two formats [15] and both have fundamental performance limits well beyond what is achievable with current converter technology. Still, despite their high promises and impressive technological potential, the DVD-A and SACD formats have both failed to gain mass-market appeal [16]. This coincides with severe problems for the music business as a whole; as the internet media revolution threatens to put both hi-res audio and the conventional recording industry out of contention [17].

From an engineering point of view, internet music shifted the prime focus of digital audio research. Since internet lines offer limited bandwidth the emphasis has been moved to *compression* and optimization of quality versus data-rate trade-offs, using sophisticated quantization and coding methods based on perceptual modelling of the auditory system [18].

Perceptual coding algorithms vary in complexity, but all are fundamentally based on the ear's *masking property*; that loud sounds overwhelm weaker sounds and render them inaudible. The hearing has both temporal, spatial and frequency based masking properties that have led to some quite sophisticated models and compression formats. This is only tangentially relevant to data converter design and the only masking property touched upon in this thesis is simultaneous or frequency masking: The hearing threshold depends greatly on the distance in frequency to a strong signal component or "masker", which is illustrated in fig.4. Consequently the spectral properties affect the severity of many distortion mechanisms. Distortion audibility will depend on the distance to maskers as well as the harmonic coherence of the distortion spectrum [19].

Modern computer based compressed audio formats are generally scalable, and with rapid increase in network and storage capacity high bandwidth transfer is gaining in popularity. Combined with advances in the sophistication of perceptual compression routines, the dynamic range and bandwidth limitations are catching up to SACD and DVD-A levels. This means that converter technology is again becoming the limiting factor of the ADDA process.

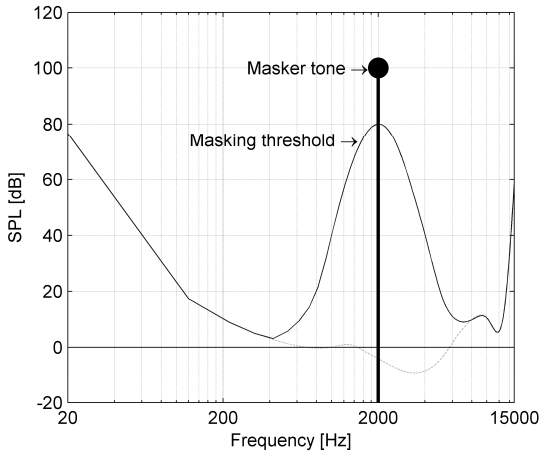


Figure 4: Conceptualization of simultaneous masking

As part of the digital audio history, growing concerns among both consumers and mastering engineers about the so-called "loudness war" [20] should also be mentioned. The omnipresence of reproduced music has led to a move from record labels to increase the nominal volume in recordings. The intention of doing so is to make a record stand out in the plethora of airwave broadcasts and marketing, since people will notice a sound quicker if it is loud. This means that the high dynamic range of modern digital formats is often not utilized. Unfortunate as it may be this is however not related to the capabilities of converter or digital format technology per se, and is thus only mentioned in the introduction for its contextual and historical relevance.

1.3 Organization of This Thesis

In the introduction, the motivation for the work has been epitomized, based on a brief review of some fundamental psychoacoustic limitations and a historical retrospect of digital audio. The next chapter will bridge this with fundamental data converter theory. It reviews the processes of sampling and quantization, as well as DA conversion and what waveform errors will typically be introduced by a DAC circuit. This chapter will also establish the case for using oversampled conversion and delta-sigma modulation in audio.

Following this, the third chapter moves on to explore the DSM and its properties. The history of delta-sigma modulation as well as principles and complications surrounding its implementation are reviewed. The concepts of stability and loop filter design are introduced, and the chapter also takes a brief look at some more recent structures and why they are used. The reader should through this gain a pragmatic understanding of delta-sigma.

The fourth chapter deals with static DAC errors and how these will limit the performance of the DA conversion process. It introduces DEM and the notation used in the fourth paper to argue for a simple estimation method to predict errors in generic DEM DACs. In addition to traditional rotation based DEM, it also explains the reasoning behind some alternative structures that have been introduced in more recent times.

The fifth chapter deals with dynamic DAC errors and how these will limit the performance of the DA conversion process. Since dynamic errors are waveform dependent, it means they will be strongly affected by the output sample sequence from the delta-sigma modulator. This sequence is generally impossible to predict analytically, but the chapter shows how its spectral properties can be used to create dynamic error estimates. This chapter has significant overlap with the contents of paper four, but was included in the monograph to make it appear more complete and coherent.

The monograph is primarily intended to provide an overview with a unified notation of the subjects touched upon in this Ph.D. project work. Having read it, the reader should be provided with the foundation necessary for a general understanding of the papers, their relevance and what their contributions constitute. The papers are themselves the main contribution; their contents having been briefly reviewed in the abstract. They are to be found in appendixes two to seven, whereas the first appendix reviews the DFT and discrete time spectral analysis of finite length signals. Such analysis is used in most converter performance evaluations, both in this work and generally, and it is therefore important to understand the properties of the DFT and the limitations and pitfalls in finite length spectral analysis.

Chapter 2

Fundamental Theory

In this chapter basic data converter theory is reviewed; it is described how data conversion works and what fundamental limitations and practical errors are inherent in ADC and DAC processing. They must be assessed in the context of sound perception as reviewed in the first chapter, forming the cognitive basis for understanding the thesis and its contents.

2.1 Sampling and Reconstruction

The fundament for digital signal processing was to a large extent made with the breakthrough discovery of the *sampling theorem*. It was implied as early as 1928, through the derivation by Harry Nyquist [21] that a system of bandwidth B could transmit independent pulse samples at a rate $2B$. Nyquist's work focused on transmission capacity and did not consider sampling and reconstruction of continuous-time signals as such. The now obvious duality of Nyquist's discovery – the theory of how any continuous-time signal can be sampled with no loss of information given a sampling frequency of at least twice its bandwidth – was first formulated by Soviet information theory pioneer Vladimir Kotelnikov in 1933[1] [22] and made known to the larger international scientific community through Claude Shannon's legendary 1948 publication "*A Theory of Communication*" [23]. Shannon formulated the theorem in "*A Theory of Communication*", and gave its proof and coined the term "sampling theorem" in his 1949 follow-up paper "*Communication in the Presence of Noise*" [24]. These two papers are generally acknowledged to be in large part the origin of modern information theory and digital signal processing, and Shannon is renowned as "the father of information theory". The sampling theorem is also often called Shannon's sampling theorem and sometimes – incorrectly – Nyquist's sampling theorem. The bandwidth limit at half the sampling frequency for any sampled signal is known – rightfully – as the Nyquist frequency.

2.1.1 Sampling

In order to enable a digital representation of a signal, samples must be taken for which to assign data values. The signal is measured at a fixed interval T_s hereafter called the sampling period. The inverse of the sampling period is known as the sampling frequency or $f_s=1/T_s$.

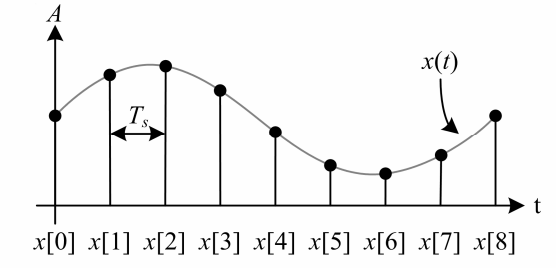


Figure 5: Sampling of a continuous-time signal

[1] Whittaker arguably gave the theorem first, implicitly [25]. History enthusiasts may enjoy the IEEE anniversary review [26].

Sampling is illustrated with a simple sinewave in fig.5. From the figure it is seen that in mathematical terms sampling is the multiplication of the input signal with a string of Dirac pulses at all integer multiples of the sampling period T_s. The mathematical description of this operation is given by:

$$x[n] = x(t) \cdot \sum_{n=-\infty}^{\infty} T_s \cdot \delta\left(t - nT_s\right). \tag{1}$$

With the Fourier transform $\mathfrak{F}$ – its definition and use assumed familiar to the reader – an expression for the sampled signal frequency spectrum $\mathfrak{F}\{x[n]\}$ can be found as a function of the continuous signal frequency spectrum $\mathfrak{F}\{x(t)\}$:

$$\begin{aligned} X_S\left(f\right) &= \mathfrak{F}\left\{x\left[n\right]\right\} \\ &= \mathfrak{F}\left\{x\left(t\right) \cdot \sum_{n=-\infty}^{\infty} T_s \cdot \delta\left(t - nT_s\right)\right\} \\ &= \sum_{n=-\infty}^{\infty} \mathfrak{F}\left\{x\left(t\right) \cdot e^{\mathrm{i}2\pi n f_s t}\right\} \\ &= \sum_{n=-\infty}^{\infty} X\left(f - nf_s\right) \ , \ \ X\left(f\right) = \mathfrak{F}\left\{x\left(t\right)\right\} \ . \end{aligned} \tag{2}$$

It is seen from this result that sampling gives a spectrum repeating around multiples of f_s as illustrated in fig.6. From both the equation and the figure it is now understood how having the sampled signal x bandlimited to below $f_s/2$ – the Nyquist frequency – is a requirement for preservation of its spectral integrity. If the repeated spectra – known as *aliases* – are removed during DA conversion, the ideal ADDA process leads to an output identical to its input.

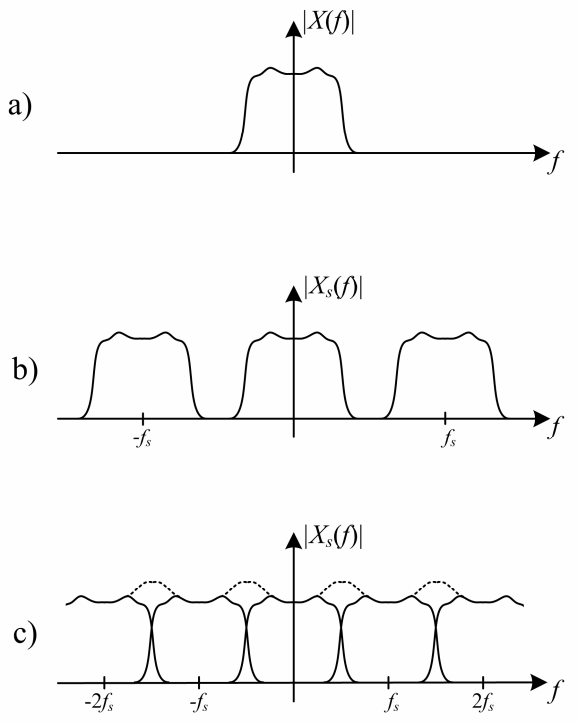


Figure 6: a) Continuous spectrum b) Sampled spectrum c) Alias distortion

If the signal bandwidth on the other hand exceeds the Nyquist frequency, or analogously that f_s violates the sampling theorem, the aliases will overlap as illustrated in fig.6c). Spectral integrity is then lost in the overlap region. In fact *any* energy content residing above the Nyquist frequency at the point of sampling will create an alias below it. It is known as *alias distortion* or just *aliasing*. From this it is given that unless the input to an ADC is limited strictly below the Nyquist frequency, alias distortion will compromise its performance. It is thus necessary to ensure that as little energy content as possible violating the sampling theorem enters the ADC. This is done by using an *antialias filter* before sampling to suppress any energy that may exist above $f_s/2$. The necessary damping of this filter is determined by the expected amount of out-of-band energy and the required level of signal integrity preservation in the baseband.

An intuitive way to understand *why* sampling produces a repetitive spectrum is to look at fig.5 and acknowledge that other high frequency sinewaves can be defined by the *exact same* sequence of samples. Thus the sample sequence contains information of many waveforms. Figure 7 shows two sinewaves giving exactly the same sample sequence. If the high frequency sinewave was the one sampled to generate this sequence, obviously in violation of the sampling theorem, reconstruction would form its low frequency alias from the samples.

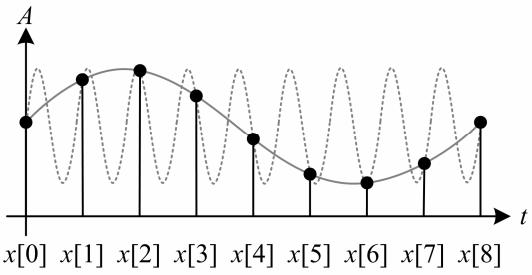


Figure 7: Sampled waveform of fig.5 and an alias

Schematically the sampling process can be seen as an AAF followed by a sampling network or an ADC. The desired input signal, filtered to conform to the sampling theorem and entering the sampler, is in this thesis denoted as $x(t)$.

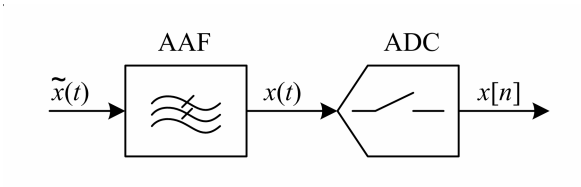


Figure 8: Conceptual ADC and AAF

Since the sampled spectrum is repetitive around f_s, and since processing the sampled signal does not necessarily imply any a priori knowledge of the sampling frequency, the sampled waveform is more conveniently expressed through its angular frequency ω defined as:

$$\omega \overset{def}{=} \frac{2\pi f}{f_s}. \tag{3}$$

Then it is given from the derivation of the sampling theorem that the frequency spectrum of the discrete sequence can be rewritten as:

$$X_s(\omega) = \sum_{n=-\infty}^{\infty} x[n] \cdot e^{-i\omega n}. \quad (4)$$

This result is the normal definition of the DTFT. It is also valid for finite length sequences as the (finite) DFT. The simulated spectra presented in this thesis and associated papers are of course of finite length and found by DFT calculation on finite sequences. The DFT may if not used carefully have incongruities due to the Gibbs phenomenon, which can be alleviated with windowing or coherent sampling as reviewed in Appendix 1. A generalization of the DTFT is given by the *z*-transform:

$$X_s(z) = \sum_{n=-\infty}^{\infty} x[n] \cdot z^{-n} \quad , z = r \cdot e^{-i\omega n}. \quad (5)$$

It is seen that the DTFT is identical to the *z*-transform for r=1, or evaluation along the unit circle in the complex plane. Although introduced for completeness, it is assumed that the reader has prior knowledge of the fundamental properties for the *z*-transform and related terms such as unit circle, poles, zeros and ROC.

2.1.2 Reconstruction

In the DAC process, the sample sequence must be transformed back to an analog continuous time waveform. Ideal reconstruction, i.e. $x_{out}(t) \equiv x_{in}(t)$, would imply removing *all* spectral content above the Nyquist frequency and retain all spectral content below it. This requires an infinitely steep reconstruction filter which is not feasible to implement. Rather, a real-life RCF is specified from how much high frequency alias energy is tolerable at the output. The RCF is typically placed outside the DAC chip as shown in fig.9. The DAC converts sample data into a continuous time waveform which is then low-pass filtered to approximate the original input. The DAC output has been given its own denotation $y(t)$. Since this thesis deals primarily with issues in DAC design, the nature of $y(t)$ is of essential interest and will be paid special attention in the theory introduction.

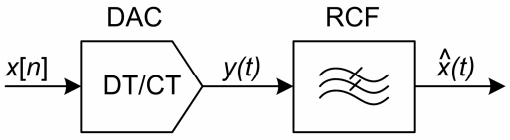


Figure 9: Conceptual DAC and RCF

The typical way of constructing $y(t)$ is to connect the output to a current or voltage proportional to the sample value and hold it over the duration of the sample period. In other words the output is defined as:

$$y(t) = x[n] \quad , nT_s \le t < (n+1)T_s. \quad (6)$$

This ensures that the output is in principle linearly proportional to the input signal *x*. The hold reconstructed waveform can also be described as the time convolution of the sample sequence and a rectangular window:

$$y(t) = \sum_{n=-\infty}^{\infty} x[n] \cdot \frac{1}{T_s} \cdot \mathrm{rect}\left(\frac{t - nT_s}{T_s} - \frac{1}{2} \right), \quad \mathrm{rect}\left(\frac{t}{T} \right) \stackrel{def}{=} \begin{cases} 1, & \frac{-T}{2} < t \le \frac{T}{2} \\ 0, & \text{otherwise} \end{cases} . \quad (7)$$

The resulting output from the DAC described in (7) is shown in fig.10 for a sinusoidal sample sequence. This is the well-known "stair-case" output waveform.

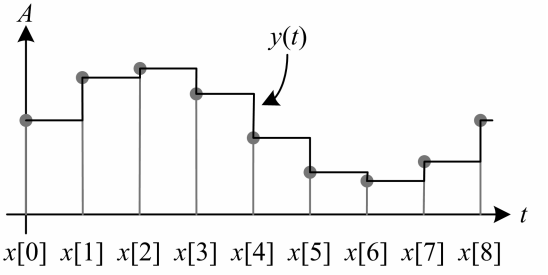


Figure 10: Output waveform from PCM DAC

The frequency spectrum of this output waveform is found by taking the Fourier transform of the time domain expression (7). Since it is known that convolution in the time domain equals multiplication in the frequency domain this is relatively simple:

$$\begin{aligned} Y(f) &= \mathfrak{F}\{y(t)\} \\ &= \mathfrak{F}\{x[n]\} \cdot \mathfrak{F}\left\{ \frac{1}{T_s} \cdot \mathrm{rect}\left(\frac{t}{T_s} - \frac{1}{2} \right) \right\} \\ &= X_s(f) \cdot \left\{ \frac{1}{T_s} \cdot \int_0^{T_s} e^{-\mathbf{i} 2\pi f t} dt \right\} \\ &= X_s(f) \cdot e^{-\mathbf{i}\pi \frac{f}{f_s}} \cdot \mathrm{sinc}\left(\frac{f}{f_s} \right) . \end{aligned} \quad (8)$$

According to usual signal processing notation, the normalized sinc-function is defined as:

$$\mathrm{sinc}(x) \stackrel{def}{=} \frac{\sin(\pi x)}{\pi x} . \quad (9)$$

Hold reconstruction in other words performs first order sinc-filtering of the sampled spectrum. This means that the aliases are suppressed somewhat, but also that there is some inband attenuation below the Nyquist frequency. This is typically compensated for at the digital side of the DAC.

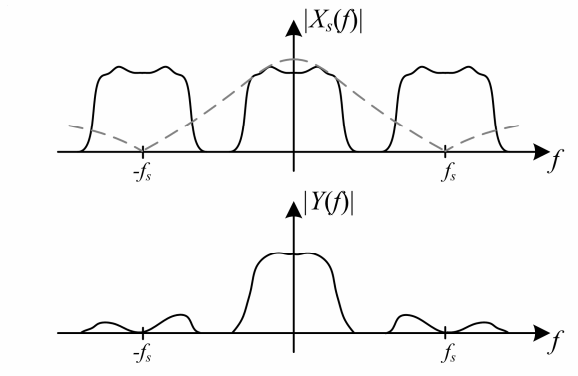


Figure 11: Hold reconstruction filtering effect

Current steering and DCT switch-cap DAC circuits, used in most audio DAC implementations, will both provide this type of waveform. The aliases are suppressed further, or analogously the "stair-case" is smoothed, through the external RCF. There also exist other types of reconstruction, some of which will be touched upon later.

2.2 Quantization

Digitization of a signal is a two-step process. After the signal is sampled, the samples must be given a data representation. The process of mapping samples to a finite set of data values is known as quantization. The most common method is scalar quantization, where each sample $x[n]$ is mapped to one in a range of values $Q(x)\in\ell$; ℓ being a scalar set that is typically integer. Another possibility is to map an input vector $\mathbf{x}=[x[1],\ldots,x[N]]$ to one in a set of output vectors $Q(\mathbf{x})\in\Re^N$, where $\Re^N$ is an N-dimensional vector space; called N-dimensional vector quantization. This thesis deals only with uniform scalar quantization which is used in practically all data converter applications.

A scalar quantizer defined by ℓ being the integer set $\{-2^{B-1}\text{-}1\ldots\text{-}1,0,1\ldots2^{B-1}\text{-}1\}$, can be realized with a B bit binary output. It is hence called a B-bit quantizer. Mapping the input to the nearest integer in ℓ can be done by rounding:

$$Q(x)=\left\lfloor \frac{x}{\Delta}+\frac{1}{2}\right\rfloor\Bigg\|_{Q(x)\in\ell}. \qquad (10)$$

This is a symmetric or mid-thread quantizer which has $M=2^B\text{-}1$ levels when it is B-bit. A B-bit quantizer can also have $M=2^B$ integer levels if it is made asymmetric. The denotation Δ is used for the *input-referred quantizer step-size*. It is shown graphically in fig.12 together with the quantization error e, which is the deviation of $Q(x)$ from x. It is seen that the error is constrained to $|e|\leq 1/2$ – or input-referred to $|e|\leq\Delta/2$ – as long as the output is within a range of 2^B. The corresponding input range $|R|\leq(2^{B-1}\text{-}1/2)\cdot\Delta$ is the input *non-overload range.*

Using Nyquist sampling and uniform scalar quantization to digitize signals is known as Linear Pulse Code Modulation and the resulting data as LPCM or just PCM samples [27]. This is the original and most direct/intuitive approach to signal digitization, but as will shortly become clear other modulation schemes can be used.

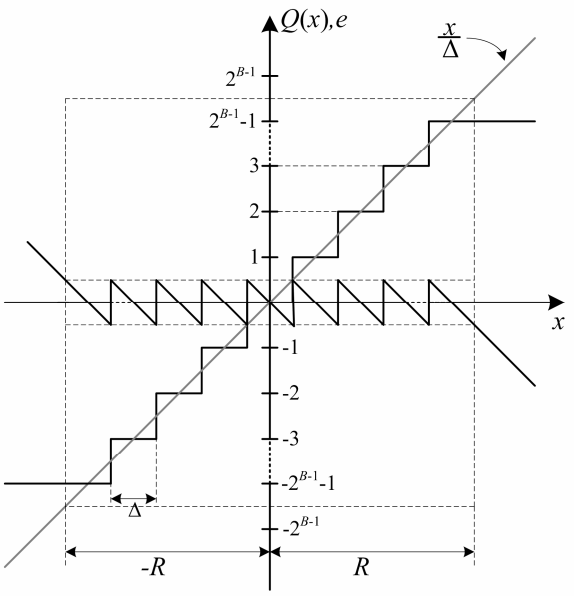


Figure 12: Uniform scalar mid-thread quantizer

The quantization error is a not a continuous function of the input signal, but has discontinuities making it difficult to analyze. In his classic paper Bennett showed that under certain conditions the quantization error can be approximated as an additive noise source, *uniformly distributed* in the range ±Δ/2 [28]. The conditions he stipulated were that the quantizer had a large input range large compared to Δ and that the input signal was active over significant parts of this range without overloading it. He proved the approximation to be asymptotically correct for a Gaussian input distribution as Δ→0, and showed through simulations it was a valid approximation for sampled and quantized sinusoids spanning over an amplitude range of many Δ. Denoting the error PDF as $f_e(e)$ it can be written as:

$$f_e(e) = \begin{cases} \dfrac{1}{\Delta} & , -\dfrac{\Delta}{2} < e \le \dfrac{\Delta}{2} \\ 0 & , \text{otherwise} \end{cases} . \tag{11}$$

From (11) the first two statistical moments of the error – i.e. the input-referred mean and variance – are given by:

$$E\left(e^m\right) = \int_{-\infty}^{\infty} e^m \cdot f_e\left(e\right) \cdot de \rightarrow \begin{cases} E\left(e\right) = 0 \\ E\left(e^2\right) = \sigma_e^2 = \dfrac{\Delta^2}{12} \end{cases} . \tag{12}$$

If the quantizer is B-bit its total input non-overload range is $2^B \cdot \Delta$. The highest level input sinewave that doesn't overload it is hence $x[n]=2^{B-1} \cdot \Delta \cdot \sin(\omega n)$, and its power $\sigma_x^2 = 2^{2B} \cdot \Delta^2/8$. The peak SQNR for sinusoid input is consequently:

$$SQNR_{\max} = 10 \cdot \log_{10}\left(\frac{\frac{2^{2B} \cdot \Delta^2}{8}}{\frac{\Delta^2}{12}}\right) = 6.02B + 1.76 \ \text{[dB]} . \tag{13}$$

This is the well known "*6dB per bit rule*" also used to calculate ENOB. Bennett in the same paper also showed that if the input signal had a smooth power spectrum, the error samples would be approximately orthogonal. Then the error autocorrelation function is given by[2]:

$$r_{ee}(k) \overset{def}{=} E\left(e[n]\cdot e[n-k]\right) \approx \begin{cases} \sigma_e^2 & , k=0 \\ 0 & , \text{otherwise} \end{cases} . \tag{14}$$

Using the Wiener-Khinchin theorem the error power spectral density is found to be:

$$S_e(\omega) = \frac{1}{2\pi}\cdot \mathfrak{F}\left\{r_{ee}(k)\right\} = \frac{\sigma_e^2}{2\pi} . \tag{15}$$

Widrow [29] extended the work of Bennett by applying sampling theory to the quantizer to find a statistical model for an *arbitrary* input PDF. This enabled Widrow to find the criteria for conditional input independence in any statistical moment of the quantization error. While the input PDF is a continuous function, the output has discrete probabilities in the value set ℓ, or input-referred in multiples of Δ. This means that the output PDF is an area sampled version of the input PDF with "sampling frequency" $\phi_q = 1/\Delta$. The probability for the output to take any discrete level is given by the cumulative input PDF within $\pm\Delta/2$ of this level.

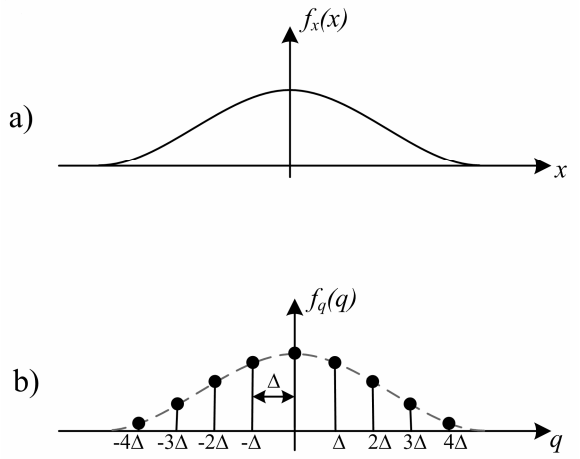


Figure 13: Quantizer input PDF (a) and output PDF (b)

For simplicity of notation, the quantizer output $Q(x)$ is denoted q in the figure and in the text from here on. The discrete PDF of the quantizer output becomes:

$$\begin{aligned} f_q(q) &= \sum_{n=-\infty}^{\infty} \delta(q-n\Delta)\cdot \int_{n\Delta-\Delta/2}^{n\Delta+\Delta/2} f_x(x)\cdot dx \\ &= \sum_{n=-\infty}^{\infty} \delta(q-n\Delta)\cdot \left[\text{rect}\left(\frac{q}{\Delta}\right) * f_x(q)\right] . \end{aligned} \tag{16}$$

The definition of the rectangular window is the same as before. Similar to the derivation the sampling theorem, Widrow took the Fourier transform of the discrete output PDF to find:

[2] This is the discrete time definition of the autocorrelation function. Bennett used continuous time analysis to show the error had approximately zero autocorrelation in two arbitrary time instants t and $t+\tau$, thus the error PSD of a sampled and quantized process would be white.

$$\begin{aligned}\Psi_q(u) &\overset{def}{=} \mathfrak{F}\{f_q(q)\} \\ &= \sum_{n=-\infty}^{\infty} \Psi_x\left(u-\frac{n}{\Delta}\right)\cdot \operatorname{sinc}(\Delta u - n)\ .\end{aligned} \tag{17}$$

The Fourier transform of a PDF is known by definition as the *characteristic function*. The CF is periodic and sinc-weighted similar to the spectrum of a signal sampled and reconstructed with hold reconstruction. To avoid PDF "aliasing", the input CF must be zero above 1/(2Δ). If so the input PDF is merely convoluted with a rectangular window of width Δ, equalling the Bennett approximation of an additive error with rectangular PDF. A large Gaussian PDF converges towards such a CF, confirming Bennett's conditions.

Widrow however went further and found a requirement for conditional independence in any statistical error moment. Any moment can by found by differentiating the CF at the origin:

$$\begin{aligned}E(q^m) &= \int_{-\infty}^{\infty} q^m f_q(q)\cdot dq = \left(\frac{\mathbf{i}}{2\pi}\right)^m \cdot \left.\frac{d^m\left(\Psi_q(u)\right)}{du^m}\right|_{u=0} \\ &= \left(\frac{\mathbf{i}}{2\pi}\right)^m \cdot \left.\frac{d^m\left(\sum_{n=-\infty}^{\infty}\Psi_x\left(u-\frac{n}{\Delta}\right)\cdot\operatorname{sinc}(\Delta u-n)\right)}{du^m}\right|_{u=0}\ .\end{aligned} \tag{18}$$

If the requirement for no PDF aliasing is fulfilled, the m^{th} output moment equals the m^{th} input moment plus a constant. But there is also a weaker condition: The assumption that:

$$\left.\frac{d^m\left(\Psi_x(u)\cdot\operatorname{sinc}(\Delta u)\right)}{du^m}\right|_{u=\frac{n}{\Delta}} = 0 \quad , n \neq 0 \ , \tag{19}$$

leads to the following simplification of (18):

$$\begin{aligned}&E(q^m) = \left(\frac{\mathbf{i}}{2\pi}\right)^m \cdot \left.\frac{d^m\left(\Psi_x(u)\cdot\operatorname{sinc}(\Delta u)\right)}{du}\right|_{u=0} \\ &\rightarrow E(q) = E(x) \\ &\rightarrow E(q^2) = E(x^2) + \frac{\Delta^2}{12}\end{aligned}\ . \tag{20}$$

It follows from (20) that the error is conditionally independent and additive in its statistical moment m if (19) is fulfilled for the m^{th} derivative. Of course this cannot be ensured with the lack of any a priori knowledge of the input statistics, but as will be seen shortly one can force this condition to hold in any given statistical moment by applying *dither*.

The reader should be aware that since both Bennett's and Widrow's methods are statistical methods, validity is limited to cases of static input PDF and they are not telling of the dynamic behaviour of the quantization noise. Many studies have been made on the dynamic characteristics of quantization noise that would make for a very extensive review. Interested readers are recommended to take a look at Gray's comprehensive survey paper [30] and its references for an overview.

2.3 Oversampling

Oversampling is a technique that has become invaluable in high resolution, low bandwidth converters. DAC oversampling is helpful in making the RCF design easier and it also gives a *processing gain* allowing re-quantization to fewer bits. One of the earliest papers on oversampled DA conversion was published in 1974 [31], and a patent was filed in 1981 [32]. Oversampling DACs have been used in most digital audio units all the way back to the Philips CD100 which had an OSR of 4.

Looking first at the ADC; sampling with a rate far higher than twice the signal bandwidth can be of benefit for several reasons. First and foremost one can use a much simpler AAF. Since the sampled spectrum is periodic in f_s, the transition band of the AAF can range from f_b to f_s-f_b, where f_b is the signal bandwidth limit. An increase in f_s in other words relaxes the requirements for the AAF by making the transition band wider, and designing a high performance AFE becomes much easier. Using the Bennett approximation for quantization noise, it is also found that the total in-band noise power – being the quantization error PSD in (15) integrated over the input Nyquist range – decreases proportionally to the OSR. If the sampling rate is increased so that $f_s=f_{s\_in}\cdot L$, the in-band quantization noise power is:

$$\hat{\sigma}_e^2 = \int_{-\pi/L}^{\pi/L} S_e(\omega)\, d\omega = \frac{\sigma_e^2}{L} \quad . \tag{21}$$

The signal-band SQNR as a function of the number of bits is consequently given by:

$$\begin{aligned} SQNR_{\max} &= 10 \cdot \log_{10}\left(\frac{\sigma_x^2}{\hat{\sigma}_e^2}\right) \\ &= 6.02B + 1.76 + 10 \cdot \log_{10}(L) \ \ [\text{dB}] \quad . \end{aligned} \tag{22}$$

It follows from (22) that for each doubling of L one can reduce B by half a bit and get the same SQNR. If L=256 four bits are saved, making ADC design simpler.

In a DAC the advantages of oversampling isn't as intuitively appreciated, but the same fundamental mechanisms apply. The space between aliases can be extended by first increasing the sample rate or zero-pad the signal and then low-pass filter it. This is the same as *interpolation* and it is shown in the time-domain as well as the frequency domain in fig.14. When the aliases are moved apart like this, the requirements for the analog RCF are greatly relaxed. In essence it means that the burden of filtering unwanted energy is moved from the analog to the digital domain. Digital filter implementations are much more flexible, much cheaper and have much higher performance than their analog counterparts.

Design of oversampling filters is a large field within DSP and is not reviewed in detail in this thesis. Unlike general purpose digital filters, oversampling filters are typically FIR filters since they can then be implemented very efficiently in a polyphase filter structure [33]. In audio applications the oversampling filter will typically be realized as several cascaded stages of halfband filters [34] or as IFIR filters [35], using a multiplier-accumulator realization [36]. For more insight in the design of oversampling filters the reader is recommended to read a textbook covering the subject, e.g. Mitra's "*Digital Signal Processing*" [37] chapter 13.

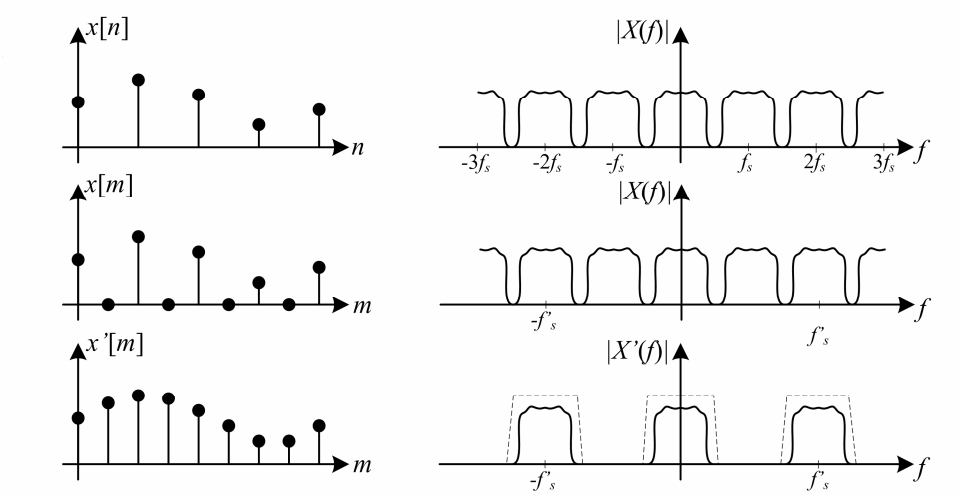


Figure 14: DAC oversampling in the time and frequency domains

With DAC oversampling the same processing gain as for the ADC will also apply to any *post-oversampling* quantization operation. This means it is possible to use a REQ to reduce the number of bits while maintaining a high effective resolution. This is shown in fig.15.

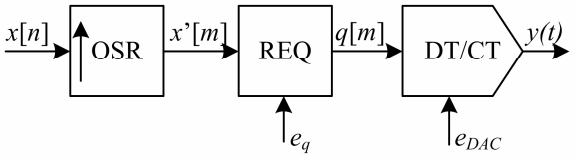


Figure 15: Oversampling DA-converter with REQ

Any noise or distortion inherent in $x[n]$ is to be regarded as part of the input signal for all succeeding processing blocks, meaning it is not reduced when oversampling. But by using oversampling, errors introduced *after* the sampling rate is increased may be spread over a larger frequency range. This is significant especially for the quantization error, which has in-band noise power as given in (21). For instance a REQ can have a 12-bit arithmetic output, but with an OSR of 256 have 16 bits effective resolution. The DAC then needs to resolve 2^{12} levels instead of 2^{16}, meaning its implementation will be much simpler. It must however be stressed that any in-band error introduced by the DAC must still be at a 16-bit level. It is only its number of elements that are reduced; the requirements for DAC in-band noise density, DAC linearity and so forth still remain the same.

2.4 Dithering

As mentioned in the section on quantization, it is possible to exploit the weaker condition of CF derivatives being zero in multiples of the quantization "frequency" to obtain conditional input independence in any error moment of choice. This is done by adding dither; a small, independent noise-source that applied to the input of the quantizer as shown in fig.16.

In the figure an additional signal v is added prior to the quantizer. Here the quantizer is a REQ used in re-quantizing DACs, meaning that the input and dither signals are generated digitally. If the input is assumed to have so high resolution compared to Δ that it can be approximated as continuous amplitude, a DAC with digital input and dither can be regarded as equivalent to an ADC with analog input and dither.

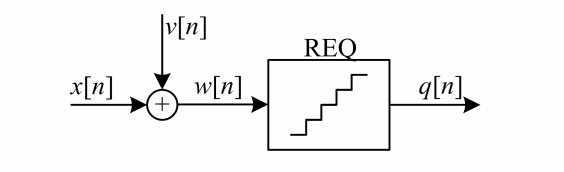


Figure 16: Dithered quantization

The first use of applied noise in quantization was seen in a 1962 publication on low-resolution image digitization [38], while the term dither was established two years later [39]. Its etymology comes from the word "didder" which means to shiver or shake with cold. The term was coined because the noise was seen to "shake up" perceptually annoying quantization error patterns. In the past, especially with regards to digital audio, there has been some dissension on the nature of dither and requirements for dithering. Results published from the research of Lipshitz, Vanderkooy and Wannamaker [40]-[42] have been very central to the development of an understanding of dither in the audio community. Their work is based on Widrow's statistical model of quantization.

Looking at fig.16, the quantizer input is now $w=x+v$. The dither signal is assumed statistically independent of the input so the PDF f_w is simply a convolution of f_v and f_x. Eq.(16) rewritten for the dithered case then becomes:

$$f_q(q) = \sum_{n=-\infty}^{\infty} \delta(q - n\Delta) \cdot \left[\text{rect}\left(\frac{q}{\Delta}\right) * f_v(q) * f_x(q) \right] . \tag{23}$$

Consequently the rewritten CF becomes:

$$\Psi_q(u) = \sum_{n=-\infty}^{\infty} \Psi_v\left(u - \frac{n}{\Delta}\right) \cdot \Psi_x\left(u - \frac{n}{\Delta}\right) \cdot \text{sinc}(\Delta u - n) . \tag{24}$$

Going back to (19), we need the m^{th} derivative of Ψ_q to be zero at all integer multiples of the quantization "frequency" for the error is to be input independent in its m^{th} statistical moment. What is now noteworthy is that if *either* of the products in (24) is zero, the whole expression becomes zero. This means that if the dither sequence is made to conform, it does not matter how the input signal behaves. Then the m^{th} output moment will regardless be given as:

$$\begin{aligned} & E(q^m) = \left(\frac{\mathbf{i}}{2\pi}\right)^m \cdot \left. \frac{d^m\left(\Psi_x(u) \cdot \Psi_v(u) \cdot \text{sinc}(\Delta u)\right)}{du^m} \right|_{u=0} \\ & \rightarrow E(q) = E(x) + E(v) \\ & \rightarrow E(q^2) = E(x^2) + E(v^2) + \frac{\Delta^2}{12} \end{aligned} . \tag{25}$$

Thus, with dither no a priori knowledge about the statistics of the input signal is necessary. All that has to be done to ensure conditional error independence, is to apply a dither signal where its given CF derivative is zero in all integer multiples of $1/\Delta$. As it turns out the sum of N independent random sources with uniform distribution in $\pm\Delta/2$ has a total CF of:

$$\Psi_v(u) = \text{sinc}^N(\Delta u) . \tag{26}$$

For this function all derivatives from 0 to N are zero in $u=k/\Delta$ for all k. The dither mean is zero and its variance is $N\Delta^2/12$. Using a single random source (N=1) is called RPDF

(rectangular PDF) or 1PDF dithering. Adding two independent sources (N=2) makes f_v triangular in ±Δ and it is therefore referred to as TPDF (triangular PDF) or 2PDF dithering. A total of N added sources – NPDF dither – will render the first N quantization error moments input independent. Studies suggest that only the first two moments – the mean and the variance – make the error audibly different (from white noise) if they are input dependent [43]. Even very coarse quantization gives no detectability of dependence in the skew, kurtosis or higher error moments. Since additional noise power from the dithering increases with N, TPDF is therefore regarded as optimal dithering in audio.

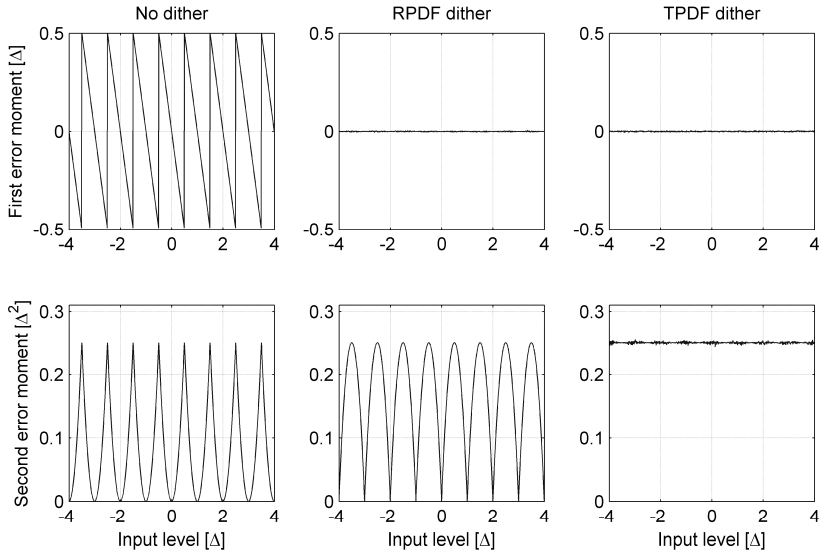


Figure 17: First two error moments as function of input level

Figure 17 shows through simulations how dither renders error moments conditionally input-independent. With RPDF dither the error mean is zero implying no distortion of the input signal. With TPDF dither both the error expectance value and the error power is constant over the input range. The average error power is increased from $\Delta^2/12$ to $\Delta^2/4$ which needs to be included in SQNR estimation (13). With TPDF dither we in other words have $SQNR_{max}$=6.02B-3.01 dB, and it is no longer just an approximation.

2.5 Delta-Sigma Modulation

Oversampling gives a nominal processing gain which enables a reduction of the number of bits while maintaining high dynamic range. Its effect in this regard is however limited; since the processing gain as mentioned is one half bit per doubling of the sampling rate. A *substantial* reduction of the number of bits – which is very desirable for complexity reasons – will require an unfeasibly high OSR.

This led to research on modulation alternatives for improvement of the processing gain. The DSM, being an extension of the delta-modulator or differential PCM encoder, was first published by Inose and Yasuda in 1962 [44]. The possibility to use it to improve processing gain in oversampled data conversion was first treated in a 1969 paper by Goodman [45], although a less known patent preceding this work exists that describes in essence the same basic principle [46].

The reader should be aware that both delta-sigma and sigma-delta are commonly used terms for the same process [47]. The causal order of the process suggests delta-sigma whereas the

modulator's functional hierarchy suggests sigma-delta. This thesis uses the original and arguably most used term; delta-sigma modulation.

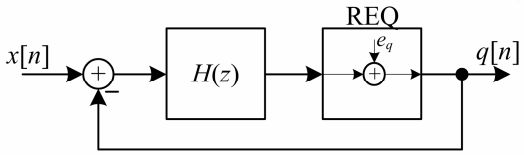


Figure 18: Basic delta-sigma modulator

The basic functionality of a DSM is depicted in fig.18. It uses filtered negative feedback compensation, causing the quantization error to be spectrally shaped. The loop filter $H(z)$ determines the spectral properties of the DSM. Using Bennett's additive noise approximation, the input-output relation of the modulator can be described by the linear sum:

$$\begin{aligned} Q(z) &= \frac{H(z)}{1+H(z)} \cdot X(z) + \frac{1}{1+H(z)} \cdot E_q(z) \\ &= STF(z) \cdot X(z) + NTF(z) \cdot E_q(z) \ . \end{aligned} \tag{27}$$

The closed loop transfer functions of the signal x and quantization error e_q have been denoted Signal Transfer Function and Noise Transfer Function respectively. It is seen that if $H(z)$ is large, $NTF(z)$ approaches zero and $STF(z)$ approaches unity meaning the signal is preserved while the quantization noise is suppressed. To achieve lowpass modulation the loop filter must be an integrator type function with high gain for low frequencies. Bandpass modulation can be realized by replacing the integrators with resonators having high gain at the frequency band of interest. The output PSD as a result of (27) is:

$$S_q(\omega) = S_x(\omega) \cdot \left|STF(\omega)\right|^2 + S_e(\omega) \cdot \left|NTF(\omega)\right|^2 \ . \tag{28}$$

Remembering the SQNR derivation (12)-(13) it is seen how an appropriate NTF can improve the processing gain since – assuming $|\mathrm{STF}(\omega)|^2 \approx 1$ – the maximum SQNR will be:

$$SQNR_{\max} \approx 10 \cdot \log_{10} \left(\frac{2^{2B}}{\frac{1}{3\pi} \cdot \int_{-\pi/L}^{\pi/L} \left|NTF(\omega)\right|^2 d\omega} \right) \ [\mathrm{dB}] \ . \tag{29}$$

An illustration is shown in fig.19, where the shaded area indicates the quantization noise falling in-band. For obvious reasons DSM is also referred to as *noise-shaping*.

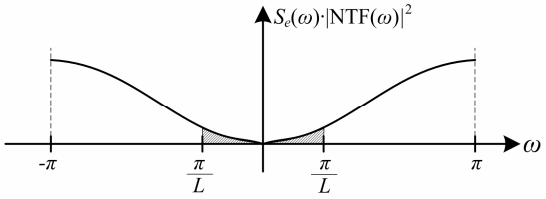


Figure 19: Illustration of DSM noise shaping

A self-evident condition for realizability is that the DSM has no delay-free loops. Thus it isn't possible to substitute $H(z)$ with a huge gain and expect noise to "disappear". The realizability condition can be formalized as $ntf[0]=1$ in the time domain or equivalently $NTF(\infty)=1$ in the z-domain. Maximum error suppression at $\omega=0$ obviously suggests all NTF zeros should be located at DC or $z=e^{i0}$. Both these conditions are fulfilled for any order N if $NTF(z)=(z-1)^N$. Often called a basic N^{th} order DSM or simply a mod$N$ in the literature, fig.20 shows processing gain in bits (according to the 6dB per bit rule) vs. OSR for $N=0$ (only oversampling) to $N=5$. It is seen that if the order is high, the processing gain is very large.

Using high order DSM with 1-bit REQ quickly became very popular in audio since a 1-bit DAC is guaranteed to have static linearity. The first audio converters were PCM converters [48], but high order 1-bit DSM quickly took over and soon reached a performance level where it in many ways outperformed the fundamental limitations of the 16-bit CD-system [49]. However while the basic functionality of a DSM is very simple, the fact that it is a non-linear feedback system creates issues not apparent when using Bennett's linear model. For instance a mod$N$ will be unstable for large input if N is higher than two and the REQ is few bits. Because of this the loop filter must be damped, causing a reduction in processing gain. The output of the DSM furthermore affects the DAC performance and its sensitivity to circuit errors. The modulator itself is also susceptible to limit cycles and noise power modulation.

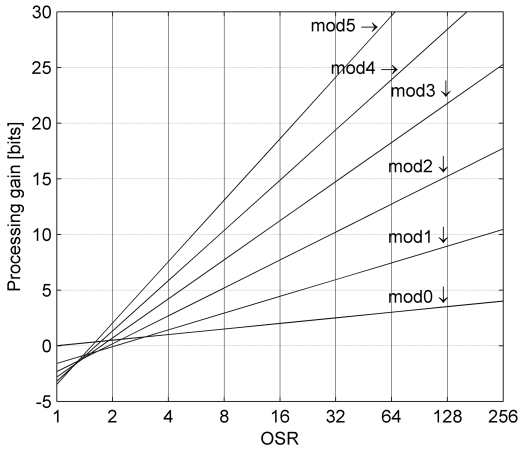


Figure 20: Processing gain of modN DSM

It is extremely complicated to do rigorous mathematical analysis of these non-linear effects and design is often based on simplified rules-of-thumb. An important part of this thesis is to present extensions of the rules-of-thumb, including a wider scope of error sources and enabling easier estimation of performance as a function of DSM design parameters.

2.6 The DAC

This section presents the DAC and gives an introduction to the errors it commonly causes. As will later be seen these errors interact with the DSM REQ and performance estimates can be given if one knows the DSM design parameters.

2.6.1 DAC topologies

Early DAC implementations – e.g. [48] – were commonly realized as resistor string DACs. An example of a resistor string DAC is shown in fig.21, where the different bits of the binary DAC input data are denoted $b_0 \ldots b_{B-1}$. Depending on whether bit b_i is one or zero, the switch it

steers is throughout the sample period connected either to ground or to the reference voltage through a corresponding resistor in a binary weighted resistor string. The output voltage is then given by:

$$\begin{aligned} V_o &= R_F \cdot V_{ref} \cdot \left(\frac{b_0}{2R} + \frac{b_1}{4R} + \frac{b_2}{8R} + \cdots \right) \\ &= \frac{R_F}{R} \cdot V_{ref} \cdot \hat{q} \ . \end{aligned} \tag{30}$$

The DAC input $\hat{q}$ is the REQ output q offset to unipolar representation, since the DAC uses positive binary values (elaborated in 2.6.2., "DAC encoding").

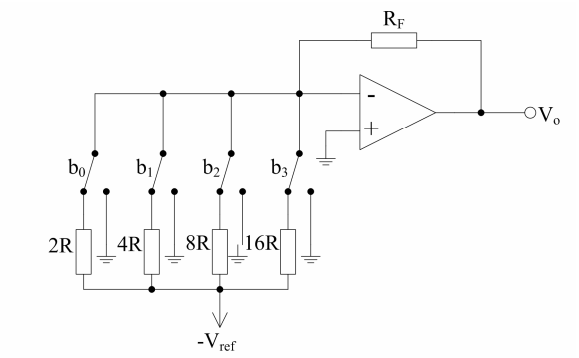


Figure 21: Resistor ladder type DAC

Use of resistor string DACs gradually lessened because technologies for IC implementation are not very suitable for large resistive devices. The resistor string will use much die area and have poor device matching. Resistor string DACs were eventually superseded by switched capacitor ("switch-cap") DACs. Switch-cap is a technique to realize resistor equivalents through charge transfer in clocked capacitors, first shown in 1977 [50]. It transfers sampled charge packets creating a resistor equivalent $R_{eq}=1/(C \cdot f_s)$, and can thus be used to implement continuous amplitude amplifiers or filters with a discrete-time transfer function $H(z)$.

In a typical switch-cap system the output of a functional stage is sampled by the next stage, meaning that only the settled output value matters. But since a DAC output is continuous-time it is very important that a switch-cap DAC settles linearly. It is possible to realize a switch-cap integrator which is insensitive to op-amp slewing and nonlinearity, called a direct charge transfer integrator. It is distinguished by the input capacitor directly depositing charge on the integrating capacitor. The DCT integrator was proposed by Bingham in 1984 [51] and a high performance DCT-based audio DAC was shown in 1991 [52]. Implementations with very high performance [53] and efficiency [54] have been seen since.

A DCT -based switch-cap DAC is shown in fig.22. In the sampling phase φ_1 it charges the sampling capacitor array depending on the input sample data, and in the hold phase φ_2 it distributes this charge directly to the integrating or hold capacitor C_h. Evaluating the charge redistribution it is found that the input-output transfer function of this DAC will be given by:

$$H_{DAC}(z) = \frac{V_o(z)}{\hat{Q}(z) \cdot V_{ref}} = \frac{\dfrac{\sum C_s}{C_h}}{\left(\dfrac{\sum C_s}{C_h} + 1 \right) - z^{-1}} \ . \tag{31}$$

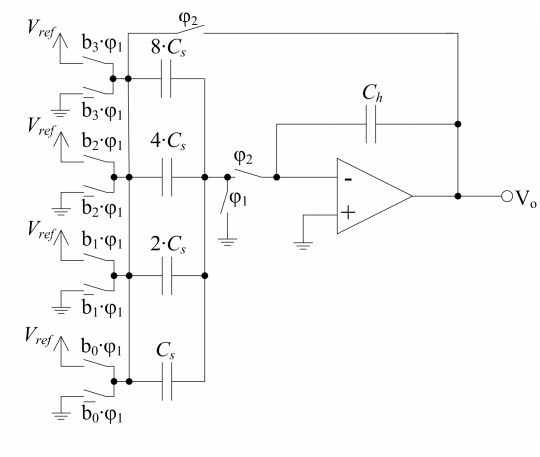


Figure 22: DCT integrator SC DAC

The low-pass function of the DCT-based DAC will be of benefit to suppress out-of-band noise from the DSM REQ. With the charge distributed passively between the capacitors, the settling is given by a linear RC time constant of the capacitors and switches, and the circuit is insensitive to op-amp slewing. Distortion is still generated from signal-dependent charge injection and signal-dependent switch resistance variation, which must be alleviated through good circuit design [53]. Still, its properties make the circuit very suitable for DAC use.

Although the DCT-based DAC still is quite popular in audio converter ICs, it has in recent years started receding. Instead it becomes more and more common for hi-res DACs to have current mode output. Then the DAC generates and holds an output *current* proportional to the input data, which is externally converted to voltage. The chief reason for doing so is that lowered supply voltages in modern IC processes reduce the headroom for SC circuits. This makes it very difficult to implement good switches and opamps, and capacitors must be big to achieve low kT/C-noise. One way to overcome these problems is to operate the DAC IC in current mode and use external I-V conversion with a dedicated high supply opamp or even a discrete transistor stage. Such an arrangement – with an opamp – is shown in fig.23. Here $I_o = \hat{q} \cdot I_{ref}$ and $V_o = I_o \cdot R_F$. In practice the external I-V is often combined with the analog RCF.

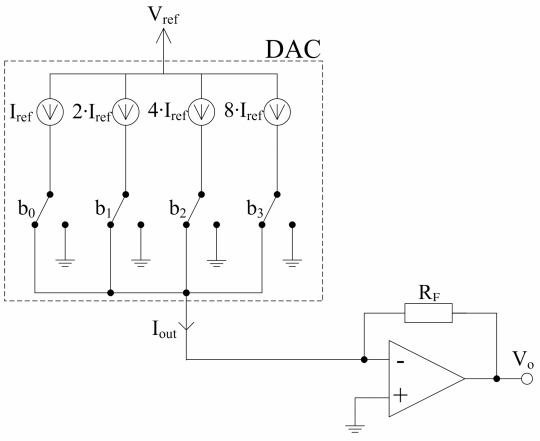


Figure 23: Current mode DAC with external I-V conversion

The idea to use steered current sources in DACs is not new [55], but it has been revived in recent times because of the development towards lower voltage IC technologies. The approach shows good potential with very high resolution having been reported [56], and it will probably be the dominant hi-res DAC design paradigm for the foreseeable future. Just like the resistor ladder DAC, the current steering DAC has no discrete time filtering of the output, and performs straightforward hold reconstruction.

2.6.2 DAC encoding

The above examples all show binary encoded DACs where the DAC elements – be it resistors, capacitors or current sources – are weighted as binary digits and fed with binary data from the (DSM) REQ. The base-two numeral system was first described as early as 800 B.C. by Indian mathematician Pingala [57] and Boole in the 19th century developed the modern concept of binary logic, the basis for all digital circuit operation [58]. Shannon was first to show automated circuits operating on Boolean logic [59], and in addition to being regarded as the father of information theory he is also widely acknowledged as the originator of digital arithmetic circuits.

The term "binary encoding" is used about non-redundant base-two arithmetic, where B bits or digits can express 2^B unique values including zero. A B-bit binary encoded DAC thus contains B elements with a 2^{B-1} size ratio between the largest and the smallest weight or digit. This is not necessarily the preferable way to implement a DAC since it is difficult to match elements with large differences in size. A much used way to get around the matching problem is thermometer encoding. The thermometer code is a redundant base-two code where every digit has unit weight. Thermometer DACs thus need 2^B-1 elements to resolve 2^B values including zero, all being equal in size.

Digital processing is usually zero mean and operating on signed binary logic, where negative numbers are represented by the two's complement of the absolute value, or $-k$ equalling 2^B-k. The first bit is then defined as the sign bit. With switching as described in 2.6.1 the DAC input must be unipolar. This means the DAC input should equal the REQ output offset by $M/2$ as shown in table 1. Thermometer encoding is offset by default.

Table 1: Binary and thermometer encoding of DAC, M=8

REQ output q	Binary q in two's compl	DAC input $\hat{q}=q+M/2$	Binary DAC input code	Therm. DAC input code
-4	100	0	000	0000000
-3	101	1	001	0000001
-2	110	2	010	0000011
-1	111	3	011	0000111
0	000	4	100	0001111
1	001	5	101	0011111
2	010	6	110	0111111
3	011	7	111	1111111

2.7 DAC errors

This section provides a review of common error sources in a typical DAC implementation. The emphasis is on the nature of the waveform distortion and how it compromises performance. For a more in-depth review of the circuit mechanisms causing DAC errors, the reader is recommended to read Wikner's thesis [60] or an appropriate textbook.

DAC errors can roughly be divided into two categories; static and dynamic errors. Static errors are errors that are time invariant whereas dynamic errors are related to the switching of elements, both leading to distortion and/or noise at the DAC output. Since the emphasis in this thesis is on current steering converters, errors are presented in the context of this DAC type and the waveform it produces. It should however be noted that the same types of errors are also present in other topologies, but then being inferred from other circuit effects (e.g. capacitor mismatch instead of transistor mismatch). Errors are normalized to the DAC input, or in other words the REQ output, thus assuming a unity quantizer step-size $\Delta=1$.

The performance is throughout assessed in terms of dynamic range measures related to noise (SNR), distortion (SFDR), or both (SNDR). Static errors are also sometimes expressed through the INL function. How INL relates to spectral performance and dynamic range is assessed in [61]; depending on its shape it may cause harmonics and degrade the SFDR or cause noise-like errors and degrade the SNR.

2.7.1 Jitter errors

The first class of DAC errors isn't really an error *in* the DAC per se, but more of an "environment variable". It has until now been assumed that T_s is always constant which in a real implementation is not the case; deviations in T_s are inevitable and referred to as jitter errors. In a typical consumer audio system, data is transferred from the digital source to the DAC through an SP-DIF connection. SP-DIF uses biphase mark encoding to multiplex data and clocking in a single coax or optical line. Band limitation in the interconnect wire will then give rise to signal dependent jitter patterns whereas transmission noise results in random jitter. The jitter behaviour of SP-DIF was analysed in an early 90s paper by Chris Dunn and Malcolm Hawksford [62].

The clock signal is recovered in the DAC through an input locked PLL oscillator. A PLL clock recovery circuit has a frequency dependent jitter transfer function, typically first order low-pass, meaning that SP-DIF transmission jitter is filtered accordingly. Important research on the JTF and the impact of jitter on conversion quality was done by the late Julian Dunn in the early 90s [63]-[64]. This was significant in establishing an understanding of jitter as an error source in digital audio, for long widely regarded as "perfect sound forever". New interface formats like IEEE1394 have been shown to be even more challenging [65].

In addition to interface jitter which is a function of the transmission and the JTF of the receiver, the receiver and DAC themselves have intrinsic clock jitter due to on-chip and on-board noise [66]. Intrinsic jitter is usually regarded as random and consisting of a white component – as a result from circuit thermal noise – and a pink component from circuit $1/f$-noise. The jitter variance, usually quantified in $(\text{ps})^2$, is typically inversely proportional to the clock frequency [67]. This means that the jitter standard deviation in ps is proportional to $1/f^{1/2}$, where in a DAC the clock frequency is typically $f=f_s=f_{s\_in}\cdot L$. The AES recently released a document for standardizing jitter terminology [68].

Analysis of jitter effects in DACs; how it results in output distortion and what kind of distortion it leads to, has been featured in several previous publications [69]-[72]. The proposed methods have often been computationally quite heavy [69]-[70], have not considered the special case of DSM [71], or have been based on experimental results [72]. The work conducted for this thesis resulted in a simple method for prediction of jitter distortion susceptibility as a function of the NTF and the number of bits in the REQ. As revealed in later chapters, the modulator output waveform significantly influences the jitter distortion susceptibility of a DSM DAC.

Figure 24 shows the error waveform of a jittered DAC with zero order hold. Whereas the ideal waveform should change value at the time instances nT_s, it in reality occurs with a jitter offset $nT_s+j(nT_s)$. This results in error pulses.

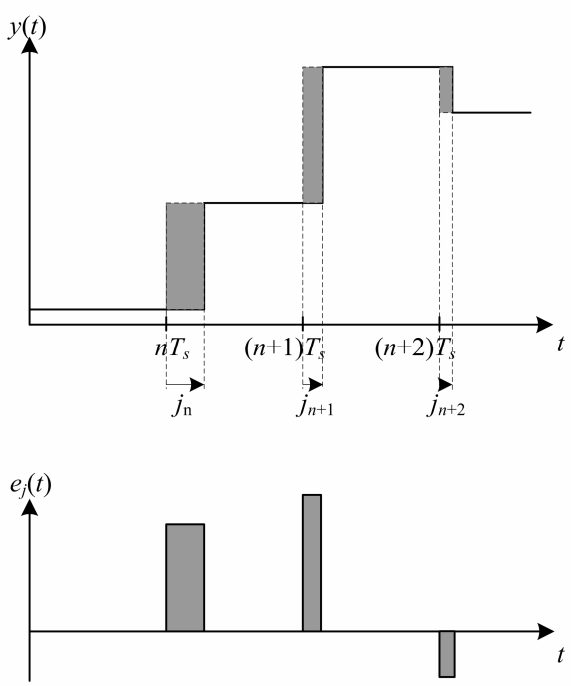


Figure 24: Jitter error in the time domain

Consider a single time instant nT_s for which the corresponding jitter is given by j_n: In an ideal DAC with hold reconstruction – i.e. where $y(t)=M/2+q[n]$, $t\in\{nT_s, (n+1)T_s\}$ – the jitter error pulse associated with *one* jittered sample is given by:

$$e_{j_n}(t) = \left[q[n]-q[n-1]\right]\cdot\frac{1}{T_s}\cdot\mathrm{rect}\left(\frac{t}{j_n}-\frac{1}{2}\right)\cdot\mathrm{sign}\left(j_n\right) . \tag{32}$$

The entire error waveform will be composed of all individual error pulses:

$$e_j\left(t\right) = \sum_{n=-\infty}^{\infty} e_{j_n}(t-nT_s) . \tag{33}$$

Consider again the single error pulse from (32): For simplicity the output step defining its amplitude is now denoted as d_n. The spectrum of this pulse is then given by:

$$\begin{aligned}
\mathrm{E}_{j_n}\left(f\right) &= \mathfrak{F}\left\{e_{j_n}\left(t\right)\right\} \\
&= \frac{d_n}{T_s}\cdot\int_{t=0}^{j_n} e^{-\mathbf{i}2\pi ft}dt\Bigg|_{j_n\geq 0} = -\frac{d_n}{T_s}\cdot\int_{t=j_n}^{0} e^{-\mathbf{i}2\pi ft}dt\Bigg|_{j_n<0} \\
&= \frac{d_n\cdot\sin\left(2\pi j_n f\right)}{2\pi fT_s}\cdot e^{-\mathbf{i}\pi j_n f} \\
&= \frac{d_n\cdot j_n}{T_s}\cdot\mathrm{sinc}\left(j_n f\right)\cdot e^{-\mathbf{i}\pi j_n f} .
\end{aligned} \tag{34}$$

The spectrum for the entire error pulse train in (33) consequently becomes:

$$\mathrm{E}_j(f) = \sum_{n=-\infty}^{\infty} E_{j_n}(f) \cdot e^{-\mathbf{i}2\pi f n T_s} \quad . \tag{35}$$

It is possible to approximate this with the DTFT by sampling the rectangular error pulse $e_{j_n}(t)$ after band limiting it with a brick-wall filter. A simulation model doing this is described in [69]. Its problem is that the computation time is very high if a long DFT and brickwall filter is used for each pulse and it thus has to be done with limited spectral resolution. A simpler approximation is to assume the jitter is very small in the frequency region of interest, so that $j_n \cdot f << 1$. Then (34) simplifies to:

$$\begin{aligned} \mathrm{E}_{j_n}(f) &= \frac{d_n \cdot j_n}{T_s} \cdot \mathrm{sinc}(j_n f) \cdot e^{-\mathbf{i}\pi j_n f} \\ &\approx \frac{d_n \cdot j_n}{T_s} \quad . \end{aligned} \tag{36}$$

This is a constant, i.e. a white spectrum. In other words the simplification assumes that $e_j(t)$ consists of Dirac pulses. The composite spectrum (35) can then be simplified correspondingly:

$$\begin{aligned} \mathrm{E}_j(f) &\approx \frac{1}{T_s} \sum_{n=-\infty}^{\infty} d_n \cdot j_n \cdot e^{-\mathbf{i}2\pi f n T_s} \\ &\approx \frac{1}{T_s} \sum_{n=-\infty}^{\infty} d_n \cdot j_n \cdot e^{-\mathbf{i}\omega n} \quad . \end{aligned} \tag{37}$$

As seen this is identical with the DTFT of a sample sequence where the sample values are given by the relative area of each error pulse. It is in other words a sampled *error area model*. For small jitter values it is quite accurate since the error area greatly dominates the distortion contribution. A previous study gives an assessment on this [73], and it can easily be calculated how close (36) approximates (34) for a given j_n.

From (37) and the convolution theorem it's apparent that the error spectrum will be the convolution of the spectra of d and j. Its PSD can generally be expressed as:

$$S_{e_j}(\omega) \approx \frac{1}{T_s^2}\left[S_d(\omega) * S_j(\omega)\right] \quad , \tag{38}$$

where – since $d_n = q[n] - q[n-1]$ for all n – the PSD of d will be:

$$S_d(\omega) = S_{\hat{q}}(\omega) \cdot \left|1 - e^{-\mathbf{i}\omega}\right|^2 \quad . \tag{39}$$

If the jitter spectrum is white the convoluted distortion spectrum will obviously also be white, meaning white jitter decreases the output SNR. If the jitter has a pink PSD the jitter noise density decreases proportionally to the distance from the signal component, making it rather benign due to the ear's frequency masking property. If the jitter sequence and input signal are both sinusoids there is multiplication of two sinusoidal terms and the trigonometric angle sum and difference identities can be used to find discrete mixing products at $\omega_x \pm \omega_j$.

In fig.25 the output spectrum is shown for each of these three cases. In real life the jitter spectrum will be a combination of both white, pink and sinusoid components. White and pink jitter noise is as mentioned typically caused by thermal and 1/*f* noise in the clock circuitry and by transmission noise. Sinusoid jitter can stem from parasitic coupling between signal lines or supply ripple, with a significant contribution also being the SP-DIF interface. As highlighted in [62], the jitter spectrum of an SP-DIF connection has lots of signal correlated sidebands.

Whereas low jitter noise is a matter of good circuit design and sufficient power for high SNR implementation, the jitter sideband distortion is more difficult to assess and control. Jitter noise is part of the DAC's intrinsic noise and thus limits its overall SNR performance. Sideband distortion on the other hand stems largely from external audio sources or from the interface, and is usually not included in a DAC specification sheet. Dunn suggested a so-called "J-test" [63] for standardizing measurements of DAC jitter immunity, which has been quite widely adopted in the audio community.

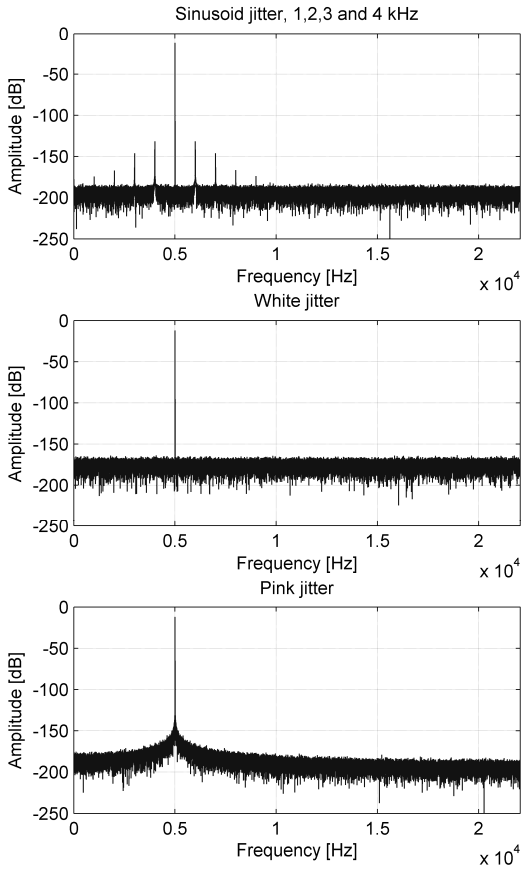


Figure 25: Jitter distortion from sinusoid, white and pink jitter

Due to the ear's frequency masking (fig.4), the audibility of jitter sidebands depends both on their magnitude and distance in frequency from the signal content. Published jitter audibility threshold estimates vary from hundreds of nanoseconds [74] to tens of picoseconds [63]. What is certain is that the jitter PSD is very important to the audibility and that HF jitter is more critical than LF jitter. Fortunately PLL-based oscillators have a low-pass JTF. It has also become increasingly popular to use asynchronous sample-rate converters which often feature JTFs with very low cut-off frequency [75]. ASRCs were originally intended to enable the connection of several digital sources with different sampling rates in one system, e.g. in a studio. But they have increasingly been incorporated in consumer audio equipment like CD-players because of jitter concerns. The thesis by Rotacher [76] provides a comprehensive review of the properties and design of asynchronous sample-rate converters.

2.7.2 Static (mismatch) errors

Static errors in a DAC are caused by physical mismatch between element weights, which always occur because of production inaccuracies. In the case of a current steering DAC there is mismatch between transistors used to realize current sources, caused by on-chip temperature deviations, threshold voltage variations and variations in the gate oxide thickness [77]. These errors are usually modelled as random stochastic variables although they may in reality be graded [78]. Error grading can be minimized through layout techniques such as common centroid, but a designer should know that random error modelling has limitations.

To generalize the notation, element weights are denoted as a set of "non-physical" variables w_i, ideally being of unity value[3]. A generalized schematic of the binary weighted DAC where $\hat{q}$ consists of bits b_0 to $b_{B\text{-}1}$ will then look like fig.26.

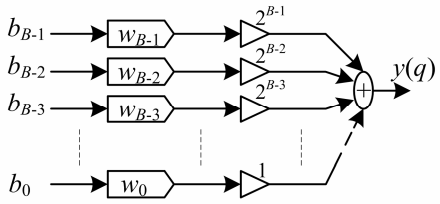


Figure 26: Generalized schematic of binary encoded DAC

The mean element weight is:

$$\overline{w} = \frac{\sum_{i=0}^{B-1} 2^i w_i}{2^B - 1} \ . \tag{40}$$

Ideally the DAC would have a perfectly linear transfer curve from 0 to 2^B-1, but because of non-ideal element weights the real one is non-linear. The deviation from linearity or INL as a function of q – derived from $\hat{q}$ =[b_0 b_1 … $b_{B\text{-}1}$] – is:

$$INL(q) = \sum_{i=0}^{B-1} 2^i b_i w_i - \sum_{i=0}^{B-1} 2^i b_i \overline{w} = \sum_{i=0}^{B-1} 2^i b_i \left(w_i - \overline{w} \right) \ . \tag{41}$$

As seen the MSB or close to MSB elements are very critical for the INL, which is the reason why binary encoding is not optimal for high resolution DACs. The relative accuracy of the largest weight must be at least $1/(2^B)$ for the DAC to have a monotonically increasing transfer function. This implies the need for transistors with a very large gate area.

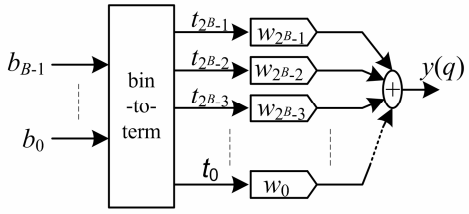


Figure 27: Generalized schematic of thermometer encoded DAC

[3] Referred to q, if referring to the input x in a re-quantizing system the ideal weight is Δ

If the DAC is thermometer encoded the mean weight of its 2^B-1 equal elements is:

$$\overline{w} = \frac{\sum_{i=0}^{2^B-1} w_i}{2^B - 1} \quad . \tag{42}$$

Thermometer encoded, i.e. with two-level representation where $t_0 \ldots t_{q-1}=1$ and $t_q \ldots t_{2^B-1}=0$, the INL becomes:

$$INL(q) = \sum_{i=0}^{\hat{q}-1} w_i - \hat{q} \cdot \overline{w} = \sum_{i=0}^{2^B-1} t_i \left(w_i - \overline{w} \right) \ . \tag{43}$$

Relative element matching is now much less critical. If the mismatch is 1% for any given DAC element, its INL contribution is 0.01 LSB. Furthermore it is guaranteed that the DAC transfer function will be monotonically increasing since the total output value always increases when more elements are connected to the output. This makes a thermometer encoded DAC, albeit having significantly higher routing complexity due to its 2^B-1 elements, often the desirable alternative.

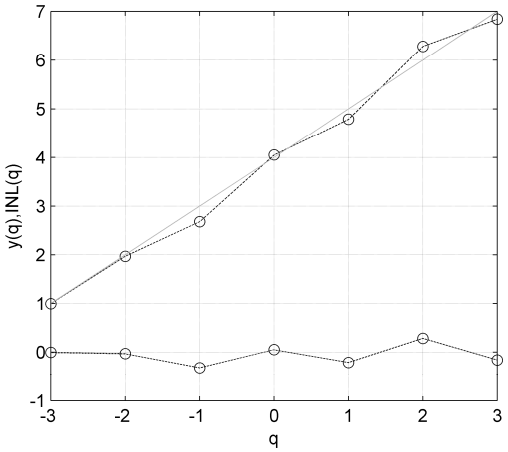


Figure 28: DAC transfer function, ideal and with INL

As will be seen later, the redundant nature of thermometer encoding can also be exploited to implement digital algorithms for mismatch-shaping, so-called dynamic element matching. DEM performs spectral shaping of mismatch errors. Although the spectral distortion as the result of a given INL curve must be found through simulations, it is later shown how to estimate it for common DEM algorithms under the assumption of random element mismatch.

2.7.3 Dynamic (switching) errors

In addition to mismatch errors which are time invariant (except if caused by temperature variations), another major source of waveform distortion is switching errors. In a current-steering DAC, switching errors can be caused by charge injection from the transistor switches as well as finite rise and fall times due to parasitic capacitances [60]. Again this thesis does not explore the circuit behaviour, but builds its analysis on generalized error waveform modelling [79].

If the DAC is thermometer encoded and all elements are identical, it can be assumed that switching on one element is associated with an on-error pulse $e_{on}(t)$ and switching off an element is associated with an off-error pulse $e_{off}(t)$. Figure 29 shows the error for one element. To derive the spectral distortion this causes will require exact knowledge about the time domain behaviour of the error waveform. This is not possible to find for any but the simplest circuit approximations and pulse simulation would still be very computationally demanding as pointed out in the jitter section. Therefore the switching error analysis just like the jitter error analysis uses error area modelling.

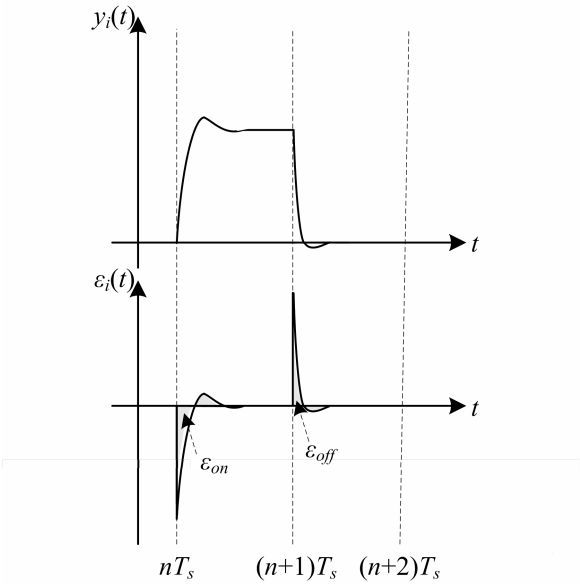


Figure 29: DAC element on and off switching and error waveform.

Error area modelling assumes that a net error area is added to or subtracted from each sample depending on how many elements are switched on or off. The mathematical analogy is again to assume that error pulses are Dirac pulses with a white spectrum and with strength given by the error area.

In a thermometer encoded DAC it is seen from table 1 that if the DAC input value increases from sample n-1 to sample n, a total of $q[n]$-$q[n$-1] elements are switched on. If the DAC input value decreases, a total of $q[n]$-$q[n$-1] elements are switched off. The error associated with the sample sequence, also called its ISI, is therefore:

$$e_{ISI}[n] \approx \begin{cases} \left(q[n]-q[n-1]\right)\cdot e_{on} = d_n \cdot e_{on} & ,d_n \geq 0 \\ \left(q[n]-q[n-1]\right)\cdot e_{off} = d_n \cdot e_{off} & ,d_n < 0 \end{cases} . \qquad (44)$$

From the ISI sequence it is relatively straightforward to calculate the waveform distortion produced at the output if q is sinusoid. As converter resolution increased throughout the 1980s, ISI became a dominant source of distortion and designers found it increasingly difficult to keep the switching errors sufficiently small. This was particularly problematic for the 1-bit DSM DACs dominant at the time, because the entire full scale value is switched each time the DSM output changes. An interesting discussion on these difficulties is found in a 1986 paper by Adams [49], and proposed solutions like return-to-zero (RZ) appeared quickly thereafter [80].

2.7.4 Finite output impedance

In addition to mismatch, another static distortion source in a current steering DAC is the finite output impedance of the current sources. For any DAC input value k there are k unit current sources coupled to the output in parallel. This means that the equivalent small signal circuit is as shown in fig.30.

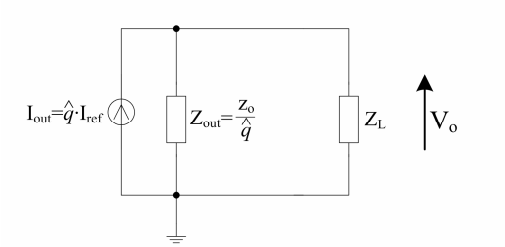


Figure 30: Equivalent small signal circuit for current steering DAC

For simplicity the output impedance is assumed purely resistive – for most cases a valid approximation [81] – and for low audio frequencies it is also assumed that the load given by the IV-converter is resistive. The output voltage then becomes:

$$V_o(q) = \frac{\hat{q}\cdot I_{ref}}{g_L + \hat{q}\cdot g_o} \quad . \tag{45}$$

In (45) g is the conductance value g=1/R. The mean LSB weight for an M-level DAC is:

$$\begin{aligned} \overline{w} &= \frac{V_o(M)}{M} \\ &= \frac{I_{ref}}{g_L + M\cdot g_o} \quad . \end{aligned} \tag{46}$$

This results in the output INL:

$$\begin{aligned} INL(q) &= \sum_{i=0}^{\hat{q}-1}\left[V_o(q) - V_o(q-1)\right] - \hat{q}\cdot\overline{w} \\ &= \frac{R_L\cdot\hat{q}\cdot(\hat{q}-M)\cdot I_{ref}}{R_o} \quad . \end{aligned} \tag{47}$$

If the current-mode DAC is differential which is usually the case in high resolution implementations, the output voltage is:

$$V_o(q) = \frac{\hat{q}\cdot I_{ref}}{g_L + \hat{q}\cdot g_o} - \frac{(M-\hat{q})\cdot I_{ref}}{g_L + (M-\hat{q})\cdot g_o} \quad . \tag{48}$$

By simple manipulation of the single-end procedure the INL is then found to be:

$$INL(q) = \frac{-g_o g_L^2\cdot\hat{q}(2\hat{q}-M)\cdot(M-\hat{q})\cdot I_{ref}}{2\cdot\left(g_L^2 + g_L g_o M + g_o^2 M\hat{q} - g_o^2\hat{q}\right)\cdot(g_L + Mg_o)} \quad . \tag{49}$$

Not surprisingly the single-ended INL is a parabola and the output SFDR is HD2 limited, whereas in the differential case the INL is anti-symmetric around the centre point and the SFDR is limited by HD3.

Figure 31 shows the INL pattern caused by finite current source output impedance for both cases. The DAC is 15-level and the *y*-axis is normalized to the LSB. The current source output resistance R_o=1MΩ and the load resistance R_L=50Ω. Given that the IV-converter represents a 50Ω load; 20-bit linearity or INL below -120dB relative to full-scale, will require current elements with an output impedance of approximately 150kΩ in the differential case and a momentous 200MΩ in the single ended case. It is thus very obvious why differential output is strongly preferred for high resolution applications.

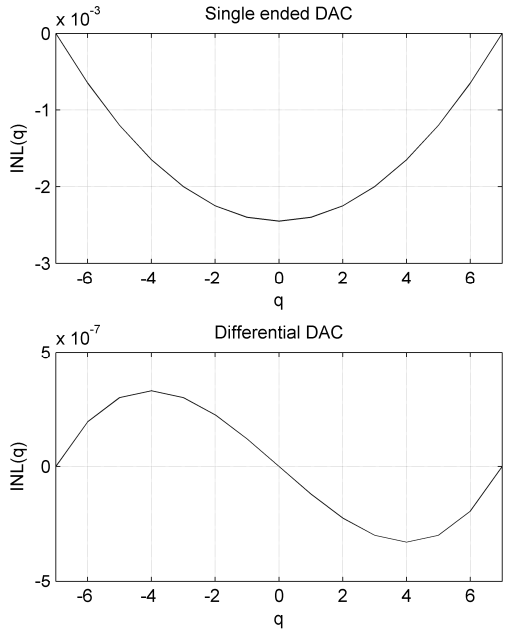


Figure 31: INL from finite output impedance

In consumer audio equipment it is very common for the full-scale output voltage to be $2V_{RMS}$. This is not a formal standard, but has established itself as a *de facto* standard. If the IV-converter is a 50Ω passive resistor the DAC output current would have to be $40mA_{RMS}$. To increase the transresistance of the IV-converter while maintaining a low load resistance, an active transresistance amplifier as shown in fig.23 must be used. For the basic circuit in fig.23 the transresistance is approximately equal to R_F if the opamp has high open loop gain, whereas the equivalent load resistance it represents is:

$$R_L = \frac{R_{in} \cdot R_F}{(A_0 + 1) \cdot R_{in} + R_F} \ . \tag{50}$$

A_0 is the opamp open loop gain and R_{in} is the opamp input resistance. Since the transresistance is approximately equal to R_F it is given that $V_o \approx I_{out} \cdot R_F$. The necessary size of I_{out} and by extension R_F is determined by the fundamental noise limitation of the IV-converter and the SNR requirement for the system.

Illustrating by example; if the IV-converter SNR requirement for $2V_{RMS}$ output is specified to 125dB@0-20kHz, its white noise voltage density must be below $8nV/Hz^{1/2}$. If the opamp's input referred noise current density is $5pA/Hz^{1/2}$, a balanced version of fig.23 with two feedback resistors and one opamp gives the following noise equation:

$$\sqrt{2 \cdot 4kTR_F + \left(5 \cdot 10^{-12} \frac{nA}{\sqrt{Hz}} \cdot R_F\right)^2} \leq 8 \frac{nA}{\sqrt{Hz}} \qquad (51)$$
$$\rightarrow R_F \leq 1\,\mathrm{k}\Omega\ .$$

From this it is found that the minimum DAC output current requirement is:

$$I_{out} \geq \frac{2V_{RMS}}{R_F} = 2\ mA_{RMS}\ . \qquad (52)$$

Note that this does not include any noise produced in the DAC itself, such as quantization noise, noise in the current sources, parasitically coupled noise, jitter noise and so on. It is therefore normal to include some headroom. Good design practice is to aim for a white noise dominated system where the IV converter's white noise specification is 3-6dB better than the total system SNR, whereas all other noise sources such as quantization noise are designed 3-6dB better than this again.

This means that if the targeted final system SNR is 120dB; the DAC output current is perhaps scaled for 125dB SNR IV-conversion, while noise contributions in the DAC – current source noise, mismatch noise, jitter noise, quantization noise and so on – are all specified to be below -130dBFS. As an example the TI DSD1792A; a high end current-mode DAC with 127dBA SNR at $2V_{RMS}$, has $2.75mA_{RMS}$ full-scale output current [82].

Chapter 3

The Delta-Sigma Modulator

This chapter takes a closer look at the delta-sigma modulator. Different architectures are described, research in stability theory is briefly reviewed and it is discussed how non-ideal behaviour affects the modulator's performance. The reader should through this gain insight in how delta-sigma modulators are designed and how things like the number of bits in the re-quantizer, the NTF and the OSR affects the performance and design conditions.

3.1 Delta Sigma Modulator Design

As seen in ch.2.5 delta-sigma modulation is in principle relatively straightforward. For convenience the basic structure and the input-output relation is repeated:

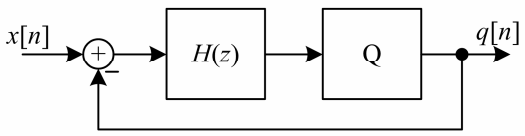


Figure 32: Basic delta-sigma modulator

$$\begin{aligned} Q(z) &= \frac{H(z)}{1+H(z)} \cdot X(z) + \frac{1}{1+H(z)} \cdot E_q(z) \\ &= STF(z) \cdot X(z) + NTF(z) \cdot E_q(z) \ . \end{aligned} \tag{53}$$

To realize mod$N$ noise shaping or $NTF(z)=(z\text{-}1)^N$ with this structure we have that:

$$\begin{aligned} H(z) &= \frac{1}{NTF(z)} - 1 = \frac{1-(z-1)^N}{(z-1)^N} \quad . \\ \rightarrow STF(z) &= 1-(z-1)^N \end{aligned} \tag{54}$$

The processing gain of this NTF is shown in fig.20 and it is seen by inspection that the STF is approximately unity for $\omega << \pi$. Although fine for most purposes, the basic modulator structure can be improved by adding a direct feed-forward path to the quantizer input as shown in fig.33. This is often referred to as a Silva-Steensgaard structure [83]. The input-output relation is now:

$$Q(z) = X(z) + \frac{1}{1+H(z)} \cdot E_q(z) \ . \tag{55}$$

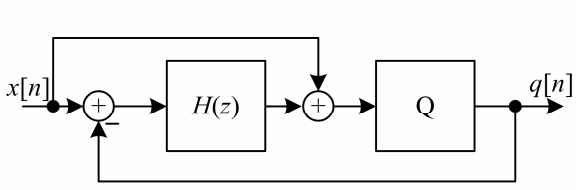


Figure 33: The Silva-Steensgaard modified DSM structure

To add this path forces a unity STF which allows the NTF to be chosen arbitrarily. Another improvement that is especially of relevance in ADC design is that the input to the loop filter $H(z)$ – which as seen from the figure is given by x-q – now only consists of shaped noise. In an ADC the loop-filter is analog and will have some degree of nonlinearity, but since the filter now doesn't process x, nonlinearity will not cause signal distortion[4].

These two DSM structures are both global feedback systems, where the output is fed back and subtracted from the input in a single node. This is not very flexible and it is entirely possible to design the STF and NTF separately by having different inputs to the loop filter for the input and feedback terms. A generalization of the basic structure with separate filter inputs is shown in fig.34.

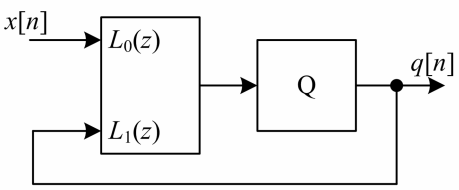


Figure 34: Generalized DSM structure

$$\begin{aligned} Q(z) &= \frac{L_0(z)}{1-L_1(z)} \cdot X(z) + \frac{1}{1-L_1(z)} \cdot E_q(z) \\ &= STF(z) \cdot X(z) + NTF(z) \cdot E_q(z) \ . \end{aligned} \tag{56}$$

It is seen that if L_0=-L_1=H, the STF and NTF are identical to the basic structure. If L_0=1-L_1 the STF is unity regardless of how L_1 and from it the NTF is designed.

A very common implementation strategy is to use a distributed feedback structure, which in its basic mod$N$ form is shown in fig.35. Looking at fig.32, the loop filter for making a mod1 NTF will be a single integrator $H(z)= I(z)=1/(z\text{-}1)$. As it turns out the DSM can be increased from mod1 to mod$N$ simply by cascading N integrators, given that the output is fed back to each integrator input. This can be verified through the generalized structure in fig.34.

The input signal goes directly through all integrators, while the feedback signal is split into N branches where each is subtracted and goes through all integrators up to the quantizer:

$$L_0(z) = \prod_{k=1}^{N} I(z) = \frac{1}{(z-1)^N} \ . \tag{57}$$

$$L_1(z) = -\sum_{k=1}^{N} I^k(z) = -I \cdot \left(\frac{I^N - 1}{I - 1} \right) \ . \tag{58}$$

[4] If the quantization error is input dependent in the first statistical moment, i.e. correlated with the input, there may be some distortion. But much less than with the basic structure.

Using (56) to find the NTF and STF it is given that:

$$NTF(z) = \frac{1}{1 - L_1(z)} = (z-1)^N \quad . \tag{59}$$

$$STF(z) = \frac{L_0(z)}{1 - L_1(z)} = z^{-N} \quad . \tag{60}$$

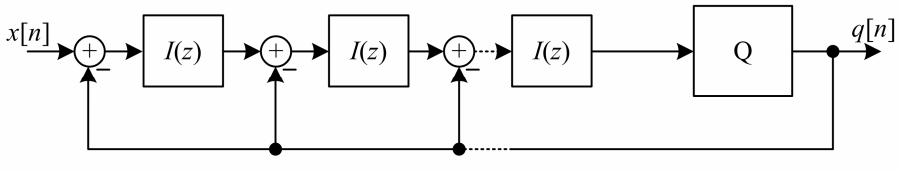


Figure 35: Basic modN distributed feedback DSM

To cascade integrators like this was actually the original idea for improving on mod1 noise shaping [31]. Analyses of second [84] and higher order [85] implementations followed. It should be noted that although $I(z)$ is here assumed to be a delaying integrator $1/(z\text{-}1)$, non-delaying integrators $I(z)=z/(z\text{-}1)$ can be used for all instances but the innermost to reduce modulator latency. A few samples latency is however normally tolerable and in a DAC this choice makes little difference. In an ADC it may be advantageous to use delaying integrators since each one can then settle independently in a switch-cap loop filter. With only delaying integrators the modulator is often called a Boser-Wooley DSM [86].

The realizability condition from ch.2.5 – that there are no delay-free loops in the system – can be formalized by writing the NTF on a form to which it must comply:

$$NTF(z) = \prod_{k=1}^{N} \frac{z - z_k}{z - p_k} \quad . \tag{61}$$

Here z_k and p_k denote the zeros and poles of the meromorphic NTF function. Until now only FIR NTFs with all zeros at DC and all poles at the origin have been considered. A problem with higher order DSMs of this kind is instability [87] and to avoid it one may have to move poles rightward, giving less peak gain around f_s at the cost of less damping around DC. Doing so is quite intuitive with distributed feedback where damping the NTF means damping the feedback terms. Maintaining STF control is done with corresponding feed-forward coefficients, giving the structure in fig.36. Such a generic structure has a high degree of NTF and STF controllability.

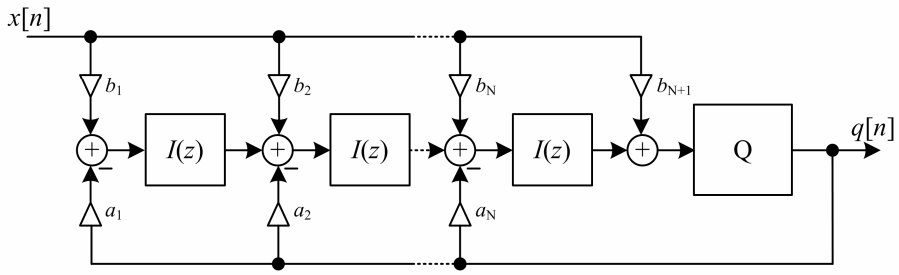


Figure 36: Generalized distributed feedback DSM

From this figure it is found that now:

$$L_0(z) = \sum_{k=1}^{N+1} b_k I^{N+1-k}(z) = \frac{b_1 + b_2(z-1) + \cdots b_{N+1}(z-1)^N}{(z-1)^N} \quad . \tag{62}$$

$$L_1(z) = -\sum_{k=1}^{N} a_k \cdot I^k(z) = -\frac{a_1 + a_2(z-1) + \cdots a_N(z-1)^{N-1}}{(z-1)^N} \quad . \tag{63}$$

And by the same algebraic manipulation as before:

$$NTF(z) = \frac{1}{1-L_1(z)} = \frac{(z-1)^N}{a_1 + a_2(z-1) + \cdots + a_N(z-1)^{N-1} + (z-1)^N} \quad . \tag{64}$$

$$STF(z) = \frac{L_0(z)}{1-L_1(z)} = \frac{b_1 + b_2(z-1) + \cdots + b_N(z-1)^N + b_{N+1}(z-1)^N}{a_1 + a_2(z-1) + \cdots + a_N(z-1)^{N-1} + (z-1)^N} \quad . \tag{65}$$

The NTF still has all its zeros at DC, but the poles are now determined by the feedback coefficients. The STF has the same poles as the NTF and its zeros are controlled by the feed-forward coefficients. If b_k=a_k and b_{N+1}=1 the zeros cancel the poles and the STF is unity. If all input branches but b_1 are removed the zeros are at infinity and the STF is $b_1/A(z)$ where $A(z)$ is the feedback polynomial. This is typically a low-pass function which can be of benefit for suppressing alias residues from the interpolation. The STF can also be designed to compensate for passband droop in the interpolation filter.

Another improvement that can be made is to replace some of the integrators with resonators to spread NTF zeros across the signal band. This can give lower total in-band noise power and improve the processing gain, especially for low OSR. It is also possible to optimize the NTF from a psychoacoustic point of view by placing zeros where the hearing is most sensitive [88]. A resonator introduces a pair of zeros at a resonance frequency $\pm\omega_r$ so each resonator must be built from two integrators. Typically a low-pass DSM has at least one NTF zero at DC, so to add one pair of non-DC zeros the modulator must be at least third order, for two pairs fifth order and so on. An example of the former is shown in fig.37. Note that at least one integrator core in each resonator must be delaying to avoid delay-free loops.

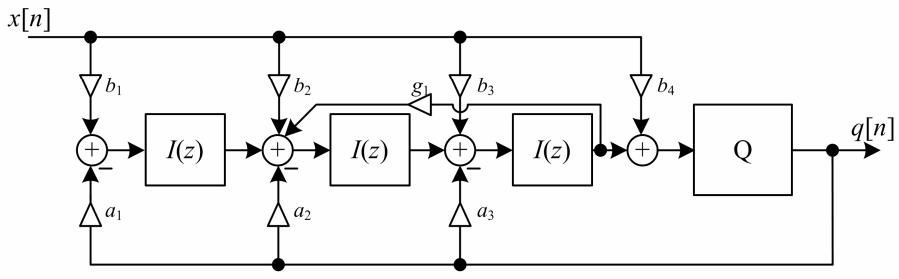


Figure 37: Distributed feedback DSM with resonator for NTF optimization

With only delaying integrators, the resonator has a local loop transfer function of:

$$R(z) = \frac{a_2 + a_3(z-1)}{z^2 - 2z + (1+g_1)} \quad . \tag{66}$$

This means that its poles, which will equal NTF zeros, are located at:

$$z_r = 1 \pm \mathbf{i}\sqrt{g_1} \rightarrow \omega_r \approx \sqrt{g_1}\Big|_{\omega_r << \pi} \quad . \tag{67}$$

It follows that if an NTF zero at e.g. 4kHz is desired and the sampling rate is 44.1kHz·64, the resonator coefficient should be g_1=0.0944. Figure 38 compares a fifth order NTF with zeros optimized for an OSR of 64 to one where all the zeros are at DC.

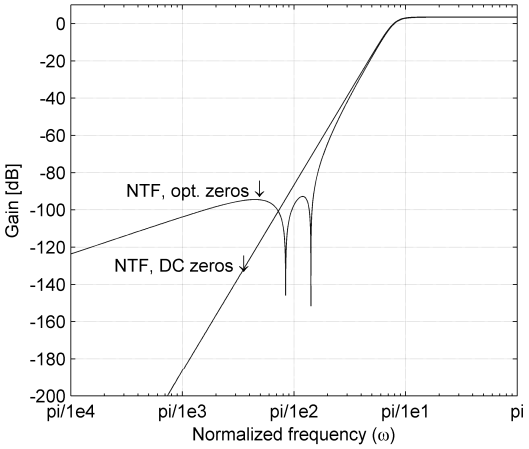


Figure 38: Optimization of NTF zeros

In addition to the described distributed feedback, another popular structure is distributed feed-forward. Then there is one global feedback path and several feed-forward branches to the quantizer input as shown in fig.39. The NTF and STF can be derived using the same procedure via the generalized structure of fig.34; this is left as an exercise for the reader. Resonators can be inserted the same way as before. The advantage with this approach is that each integrator sits in a local Silva-Steensgard loop, meaning no integrator inputs contain any signal component. As mentioned this gives linearity benefits in ADC implementations. A fifth order FF-DSM is Sony's recommendation to use in ADCs for the SACD-format [89].

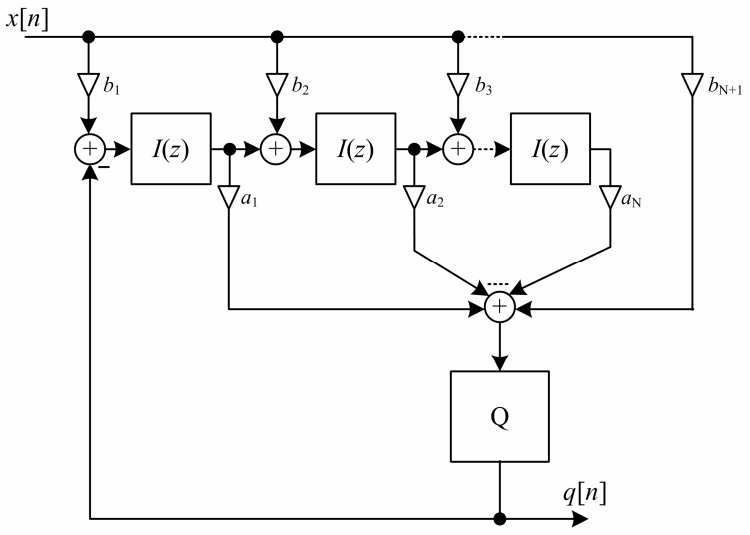


Figure 39: Distributed feed-forward DSM structure

3.2 Alternative Delta-Sigma Structures

From the basic topologies reviewed, many modified or slightly different modulators have been published where some tricks are typically used to improve a certain design parameter. To show all these would be much too comprehensive for this text, but a few of the most significant are introduced; namely the error feedback modulator, the multi-stage or MASH modulator and the Trellis modulator. All these have been shown in audio applications.

Beginning with the error-feedback modulator; this is a simple and seemingly very attractive structure that is unfortunately unsuitable for ADCs but frequently used in DACs [90]-[91]. It is shown in fig.40. The reason that it is not suitable for ADC use is apparent, since there is an analog subtraction and a loop filter in the feedback path with no error suppression from it to the final output. In a distributed output feedback DSM only the first integrator is without error suppression.

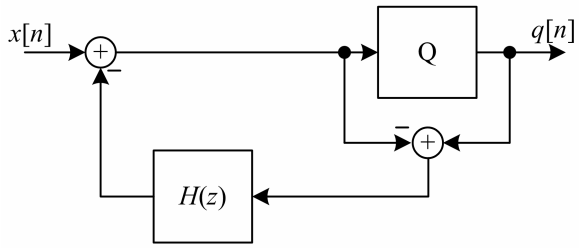


Figure 40: The error-feedback DSM structure

$$\begin{aligned} Q(z) &= X(z) + (1 - H(z)) \cdot E_q(z) \\ &= X(z) + NTF(z) \cdot E_q(z) \ . \end{aligned} \tag{68}$$

In a DAC this structure makes it simple to realize the loop filter since it is only 1-*NTF*(*z*). Especially if the modulator can handle a basic mod*N* NTF it is simple to implement *H*(*z*) since it then is a FIR filter. Unlike reported in [91], the stability constraints (more about this in ch.3.3) are the same as for a regular DSM with identical NTF and unity STF.

Another frequently used alternative is multi-stage or MASH noise shaping. MASH was first introduced in 1986 as a way to obtain mod2 and later mod3 DSMs using single integrator loops [92]-[93]. Generalized MASH was analyzed in a 1989 publication [94]. A MASH is made up of cascaded sub-modulators, and is often described as an $o_1 o_2 \ldots o_M$ MASH where o_k is the order of modulator k in the cascade. Many applications use a 1-0 MASH (often called a Leslie-Singh modulator), but in audio it is more common to use a 2-1-1 MASH. A two-stage MASH example is shown in fig.41, with its input-output relation given in (69).

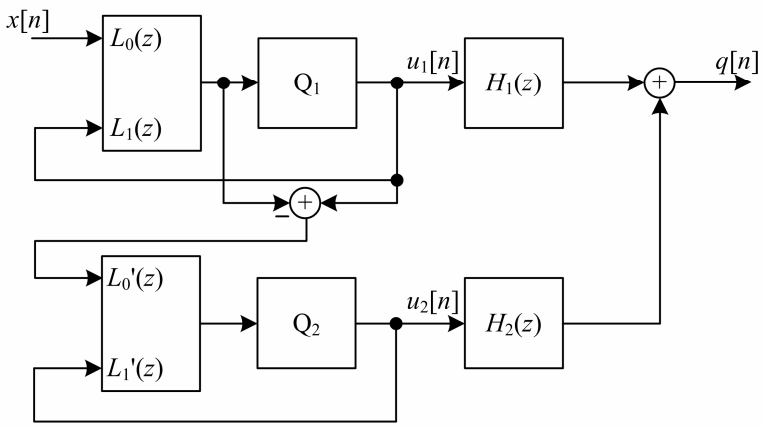


Figure 41: A two-stage MASH modulator

$$\begin{aligned} Q(z) &= H_1(z)\cdot U_1(z) + H_1(z)\cdot U_2(z) \\ &= H_1(z)\cdot\left(STF_1(z)\cdot X(z) + NTF_1(z)\cdot E_{q1}(z)\right) \\ &+ H_2(z)\cdot\left(STF_2(z)\cdot E_{q1}(z) + NTF_2(z)\cdot E_{q2}(z)\right) . \end{aligned} \tag{69}$$

The first stage quantization error will be cancelled if fulfilling the condition:

$$H_1(z)\cdot NTF_1(z) - H_2(z)\cdot STF_2(z) = 0 . \tag{70}$$

(70) is fulfilled if $H_1(z)=STF_2(z)$ and $H_2(z)=NTF_1(z)$. The output is then given by:

$$Q(z) = STF_1(z)\cdot STF_2(z)\cdot X(z) + NTF_1(z)\cdot NTF_2(z)\cdot E_{q2}(z) . \tag{71}$$

It is seen that if both the first and second stages are mod2 the total NTF will be mod4, or more generally $o_{MASH}=\Sigma o_k$. At the same time loop stability is determined by the low order sub-modulators, which is greatly advantageous. For DA conversion the disadvantage with MASH is that it cannot possibly be used to realize a two-level DAC output, since filtering and recombination of the sub-modulator output terms will produce a multi-level signal. In MASH ADCs this is not a problem, but there may be some leakage of e_{q1} since the analog modulator loop can not be made to match exactly the digital post-filter function. Still the MASH can be a useful structure for both.

The last modulator structure that is looked at in this section is the Trellis noise shaping modulator. TNSM is a look-ahead modulator scheme used to improve 1-bit noise-shaped encoding. The basic principle for look-ahead modulation is based on the “ultimate modulator” shown in fig.42, and TNSM was first introduced in a 2002 publication by Kato [95].

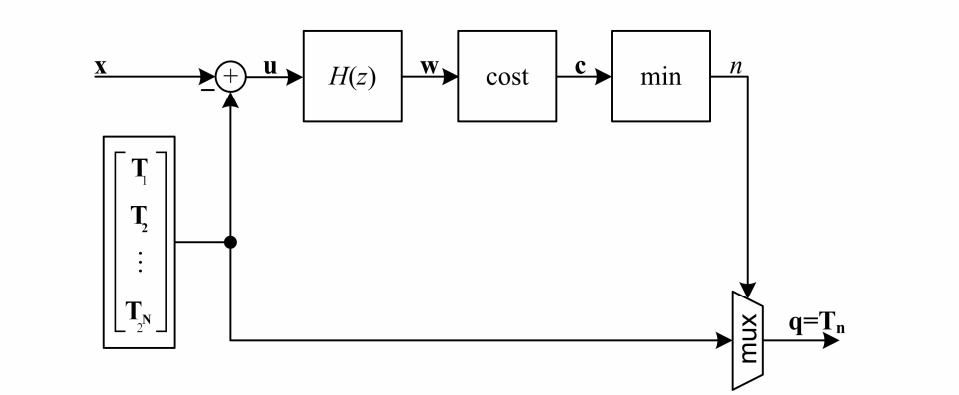


Figure 42: Principle for the "ultimate modulator"

The complete input data sequence can be written as a vector **x**. In an "ultimate" scenario, **x** should be compared to *every possible permutation* of an equally long binary output vector and the one producing the *smallest total error* should be chosen as the output. Since in-band noise is sought to be minimized, the error evaluation is weighted by $H(z)$ meaning a noise shaping function $NTF(z)=1/H(z)$ is imposed. Finding the smallest error is formalized as minimizing an error cost function, typically the MSE. The single permutation giving minimum cost is the "ultimate output sequence". Obviously this isn't feasible to implement as **x** could be infinitely long and an "ultimate modulator" would need infinite memory and storage.

Consider a single sample instant: A binary "ultimate modulator" has the choice of setting $q[n]$='0' or $q[n]$='1'. In the next sample instant it has two options of setting $q[n+1]$='0' or $q[n+1]$='1' for each $q[n]$. In other words there are four possible permutations or *candidate paths* '00', '01', '10' and '11' from n to $n+1$. There are eight candidate paths from n to $n+2$, sixteen to from n to $n+3$ and so on. The TNSM uses a variation of the Viterbi algorithm [96] to keep the number of candidate paths constrained.

For a sub-sequence of length L there are 2^L candidate paths. In the sample instant following it, either '0' or '1' can be added bringing the number up to 2^{L+1}. But if adding either '0' or '1' is determined depending on what gives lowest cost *with the existing candidates*, half the candidates can be discarded and there are only 2^L candidates left for the next sample instant as well. Saving these and accumulating the cost function, the procedure can be reiterated for minimum accumulated cost at every sample instant, generating a *candidate trellis* of width 2^L. Usually L is referred to as the *trellis order*.

If this procedure runs long enough the paths will converge, meaning that all candidates in the candidate trellis at $n=k$ are likely to have originated from *the same output* for $n=k-D$ where D is a large integer. The output can thus be unambiguously determined from backtracking of the trellis by D samples.

Figure 43 shows a general block diagram of a TNSM. The 2^L processing units are used to determine the accumulated baseband error cost from adding '0' as well as '1' to each candidate of the L^{th} order trellis. The cost metrics are then sent to a trellis generator that determines which to choose, discards half the candidates and advances the trellis one step. The new trellis "layer" must be fed back to update the filter states according to the choice made. A total of D trellis layers are stored in the trellis register. Since the paths converge the output generator can create an output sequence unambiguously by backtracking though the trellis register.

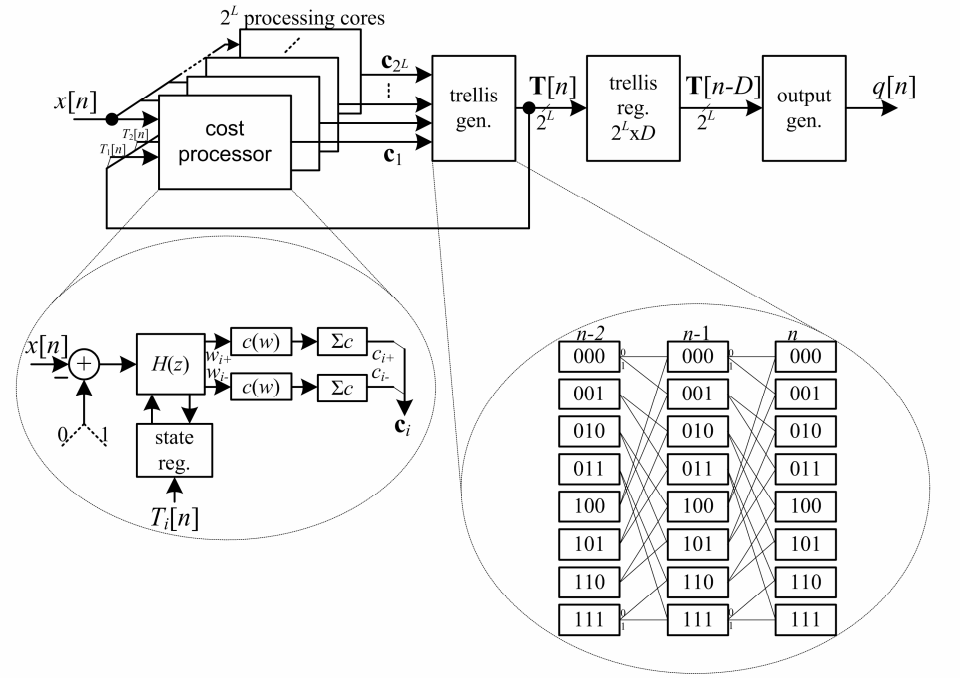


Figure 43: Trellis noise shaping modulator

In TNSM publications it is typically suggested for the trellis order L to be between two and four. The higher the better, but requirements for storage as well as the number of computations doubles for each increment of L meaning that the trellis order is limited by complexity. The backtracking depth D is typically recommended to be a few thousand samples. The cost function is usually the cumulative MSE or:

$$c_i[n] = \sum_{k=0}^{n} \left| w_i(k) \right|^2 \quad . \tag{72}$$

Here $w_i(k)$ is the filter output of processing block i at time instant k. Because of the high complexity, recent research has revolved around further path reduction while maintaining the desirable properties of "ultimate" modulators [97]. Alternative algorithms for look-ahead modulation have also been shown [98]. The advantages of the TNSM are probably better understood after reading the next sections on the non-idealities in a regular DSM. Because of a more global error optimization the TNSM is much less tonal, it is less susceptible to noise power modulation and – perhaps most importantly – it is much more stable. Whereas a traditional high order 1-bit DSM is typically limited to -6dBFS or less stable input range, a TNSM with the same NTF may work up to -2dBFS [95]. Furthermore it loses track gently, rather than having the catastrophic instability behaviour of a regular DSM.

3.3 Stability

As mentioned in ch.2.5 a mod$N$ NTF will not yield a stable DSM for high N. From this it is understood that ensuring BIBO stability in the NTF is not sufficient to know if the modulator is stable. Modelling the quantizer as an additive noise source does not take into account the fact that it is in reality a nonlinear unit with limited input range. For example one can envision a situation where the input signal x is so large that the input to the first integrator is always positive; in this case the integrator output steadily increases without bond and drives the quantizer into overload. The output then loses track of the input and typically the modulator

starts to oscillate. Figure 44 shows an example of modulator instability. For the first few samples the output tracks the input and performs noise-shaped 7-level quantization, but when the input gets too large the quantizer starts overloading and the system goes into oscillation.

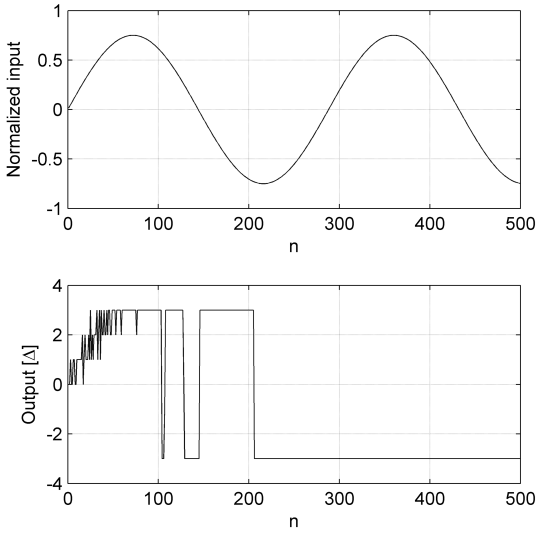


Figure 44: Example of instability in high order DSM

Analysis of instability in non-linear feedback systems is extremely difficult and for high order DSMs no analytical method to find absolute stability constraints exists. It is clear that quantizer overload is not a sufficient proof of instability, since for instance a 1-bit DSM REQ operates in overload for all non-zero input signals. Therefore design constraints are determined empirically and the designer must use extensive simulations to ensure stability in system implementations.

Some practical rules-of-thumb and non-rigorous mathematical methods have been published that provide a starting point, the most famous of these probably being Lee's Rule [99]. It formulates the following NTF constraint:

$$\left\| NTF\left(\omega\right) \right\|_{\infty} \leq 1.5 \ . \tag{73}$$

Lee found through extensive simulation work that if this constraint is met, the 1-bit DSM will most likely be stable for input up to ±0.5 normalized amplitude or -6dBFS. It must however be noted that there is no mathematical proof for this, and stability must still be validated by the designer. Some sort of automated reset function should also be included in case of instability [100]. In a multi-bit DSM the NTF can be more aggressive since the quantizer non-overload range is bigger compared to the quantization step and error. It has e.g. been suggested to use the restriction $\|NTF(\omega)\|_{\infty} \leq 3.5$ in 3-bit modulators [101]. Alternatively one can keep the NTF conservative but allow a bigger input range. This depends on whether quantization noise dominates the noise budget and it is a trade-off that should be done early in the design process. If the stable input range relative to full scale output is a value $|A_{max}|<1$, typically $0.5<|A_{max}|<1$, the maximum stable SQNR is found by modifying (29) accordingly:

$$SQNR_{\max} \approx 10 \cdot \log_{10} \left(\frac{A_{\max}^2 \cdot 2^{2B}}{\frac{1}{3\pi} \cdot \int_{-\pi/L}^{\pi/L} \left| NTF(\omega) \right|^2 d\omega} \right) \text{ [dB]} \ . \tag{74}$$

Another method that is much used in addition to Lee's Rule is the Root Locus method [102]-[103], also developed for 1-bit DSM REQs. A 1-bit quantizer can be seen as a gain element, where the gain g is inversely proportional to the input amplitude as shown in fig.45. The linear DSM model can then be modified accordingly, also shown in the same figure.

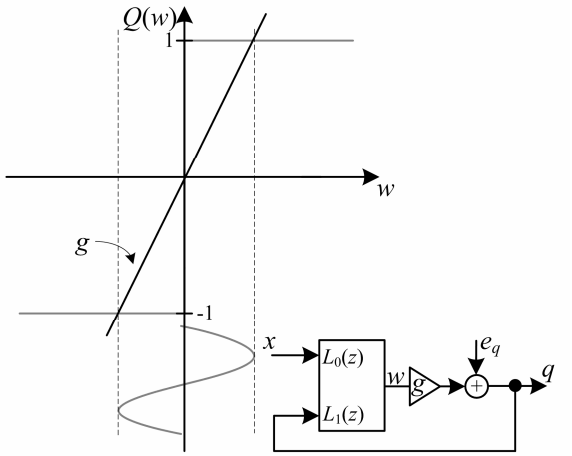


Figure 45: Modified linear DSM model used in Root Locus method

The modified NTF becomes:

$$NTF_g(z) = \frac{NTF_1(z)}{g + (1-g) \cdot NTF_1(z)} \ . \tag{75}$$

A simple check of stability can then be made by sweeping g from 0 to 1 and see if the NTF remains BIBO-stable over the whole range, or in other words check if the poles (roots) stay inside the unit circle for all g. For small g – or in other words large input – one may find that the roots move outside the unit circle, which will be an indication of instability.

Figure 46 shows the simulated processing gain for modulators designed to be stable with 1-bit quantization according to the above methods. Compared to fig.20 it is seen that the processing gain is severely restricted, especially for low OSR. With multi-bit quantization the choice of NTF is less restricted, and with more than four bits or so the processing gain can be very close to that of fig.20. Even though multi-bit quantization has now become very popular also in high OSR audio converters, it was the desire for low OSR and high bandwidth delta-sigma conversion that really drove the development of multi-bit DSM.

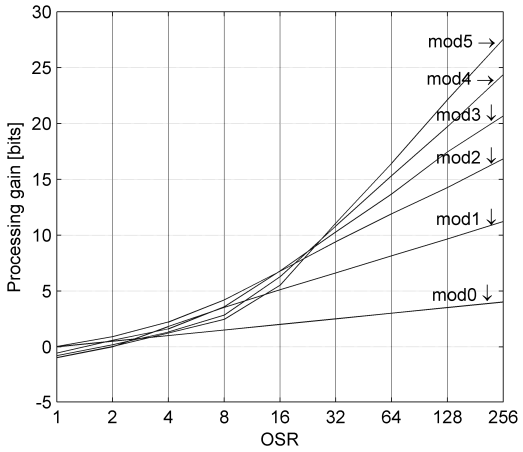


Figure 46: Processing gain with 1-bit stable DSM

The mentioned methods for stability analysis have in common that they are simple, backed by empirical results [104] but also that they lack a rigorous mathematical basis. Many attempts have been made to develop a mathematical framework for better analysis of higher order delta-sigma modulators. The quasi-linear "describing function method" was used in an early publication by Ardalan [105], whereas Hein [106] and Wang [107] pursued approaches based on geometric analysis. Several publications deal with efforts to build a framework based on non-linear dynamics; first used in DSM analysis by Freely [107] and given a comprehensive treatment in the thesis of Risbo [109]. A quite recent paper by Reiss [110] provides a historical review of non-linear DSM analysis and some general assessments of the road ahead. Reiss also presented an intriguing paper at the 124th AES Convention [111], in which parallel decomposition of the loop filter was used to break the DSM down into a sum of first order functions, mutually dependent only through the quantizer function. This provided very promising simulation results in support of a framework for stability analysis of general higher order 1-bit modulators.

Although they are significant for developing a theoretical foundation, practical application of most of the above methods is problematic due to them being very difficult to use and typically only shown with strong limitations on input conditions, initial conditions and modulator designs. Schreier et al. published significant results by estimating stability bounds based on invariant sets in the DSM state space [112], and developing computer code for how to find them [113]. But this method too is neither rigorous nor analytical.

3.4 Cyclic Behaviour, Tones and Noise Power Modulation

An expression for the processing gain of a DSM was found in ch.2.5 using the linear quantizer model. Typically when designing DSM-based converters, this method is used to choose an appropriate OSR and NTF for a target SQNR. The linear model does however hide some unfortunate effects also present during stable operation. The most serious one is perhaps cyclic behaviour. Cyclic behaviour can be understood through a simple example; a 1-bit mod1 with a rational DC-input. The output from a mod1 with a single delaying integrator is in the time domain given by:

$$q[n] = Q\left(\sum_{k=1}^{n} \left(x[k-1] - q[k-1] \right) \right) . \tag{76}$$

This can be rewritten to:

$$q[n] = Q\left(\sum_{k=1}^{n} x[k-1] \right) - Q\left(\sum_{k=1}^{n} q[k-1] \right) . \tag{77}$$

Consider a binary quantizer and an input sequence $x[n]=⅓\Delta$ for all n. Then $q[n]$ will be given as the sequence {1,1,-1,1,1,-1,1,1,-1,1,1,-1...}. Not surprisingly the output mean is ⅓ since mod1 has an NTF zero at DC. But it is seen that the bit pattern repeats and the output energy is concentrated in $f_s/3$. This is known as a *limit cycle*. Repetitive output patterns or limit cycles in mod1 was first described mathematically by Candy [114], while Friedman [115] extended the analysis to describe limit cycles in mod2. For high order DSMs it is much more difficult to find limit cycles but it has been proven that they exist [116]-[117].

A problem caused by cyclic behaviour is *idle-tones*. Whereas a limit-cycle as such describes the repetitive output pattern occurring under strictly defined state conditions, an idle-tone is a discrete component appearing in the noise spectrum during normal operation *because of* cyclic behaviour [116]. This should be kept in mind although a lot of literature does not distinguish between the theoretical limit cycle and the practical idle-tone. Idle-tones occur when the input is idle, i.e. DC. Figure 47 shows the output spectrum of a fifth order binary DSM with optimized NTF zeros and rational DC input stimuli. It is seen to have some clearly visible in-band idle-tones.

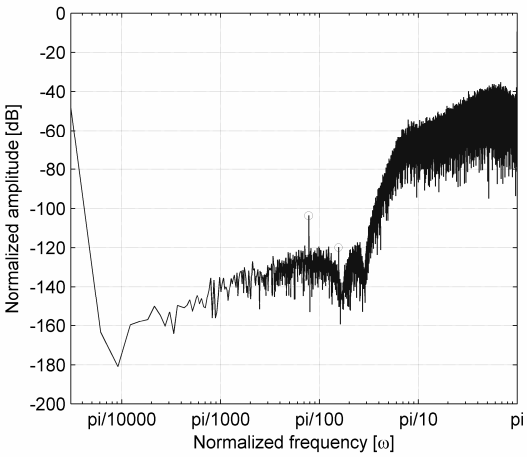


Figure 47: Output spectrum from fifth order DSM with rational DC input

Tones in the output spectrum inferred from cyclic behaviour also occur when the input is active. This has been shown for simple sinusoid input signals and in the literature it is referred to as modulator harmonic distortion [114], [118].

Since the ear is generally much more sensitive to discrete tones than to noise, cyclic behaviour is highly undesirable. It can be reduced or avoided through dithering: If the error is uncorrelated with the input the DSM will be tone-free. This is as known from ch.2.4 achieved with full *N*PDF dither of any *N*. Lesser dither weakens the correlation without removing it, thus reducing tones without eliminating them fully. Full dither is achievable with many levels in the REQ, while for a few-level or 1-bit DSM dithering even less than this will significantly reduce the stable input range. Alternative techniques have therefore been proposed like making the DSM chaotic [119]. Chaotic operation is achieved if one or more NTF zeros are

moved outside the unit circle, by modifying one or more integrators so that $\hat{I}(z)=1/(z-\alpha)$ where $\alpha>1$. The thesis of Risbo [109] investigates chaotic modulators. It is also possible to use dynamic dithering where the dither level is inversely proportional to the input level [120]. Then the dithering is strong for weak input and vice versa. As such it makes use of the ear's masking property since a loud signal will mask tones.

Another potential problem is noise-power modulation, already introduced in the chapter on quantization and dithering. It has often been argued that the DSM is a self-dithering system since quantization noise is fed back into the modulator from the output. However Wannamaker proved in his thesis [43] that since the fed back error is not input independent in any other moment than the dither forces it to be; the m^{th} derivative of the input and dither *joint PDF* has be zero in all multiples of the quantization "frequency" if the m^{th} error moment is to be input-independent:

$$\left.\frac{d^m\left(\Psi_{v,x}(u)\cdot\operatorname{sinc}(\Delta u)\right)}{du^m}\right|_{u=\frac{n}{\Delta}}=0\ ,\ n\neq 0\ . \tag{78}$$

$$\Psi_{v,x}\overset{def}{=}\mathfrak{F}\left\{f_{v,x}(v,x)\right\}\ . \tag{79}$$

With no a-priori knowledge of the input statistics this is only ensured if:

$$\left.\frac{d^m\left(\Psi_{v}(u)\right)\cdot\operatorname{sinc}(\Delta u)}{du^m}\right|_{u=\frac{n}{\Delta}}=0\ ,\ n\neq 0\ . \tag{80}$$

This is the *exact same* requirement as for the non-modulating dithered quantizer in 2.4. What it means is that to guarantee error moments 1 to m to be input-independent, a DSM REQ needs m^{th} order dithering just like an ordinary REQ. The fundamental dither requirement does not change. This postulate led to some dissention and heated debate between those claiming the DSM to be self-dithering and the purveyors of Widrow's statistical analysis [42]. It is clear that if the REQ is situated in a high order DSM loop, the input-dependency of the error is quite weak and resulting noise-power modulation quite low, just like tones are low and the distortion is low. An investigation of *practical levels of in-band noise power modulation* in several DSMs was featured in the first paper (Appendix 2), which was intended to provide a pragmatic context to this discourse. Further investigations are also featured in the recent thesis by Campbell [122].

It should be noted that the method of chapters 2.4 and 3.4 is not directly applicable to the 1-bit quantizer which has traditionally been used in audio DSM converters. Assuming the 1-bit quantizer takes the sign of the input – still denoting the quantization step Δ – its output has two discrete probabilities as shown in fig.48.

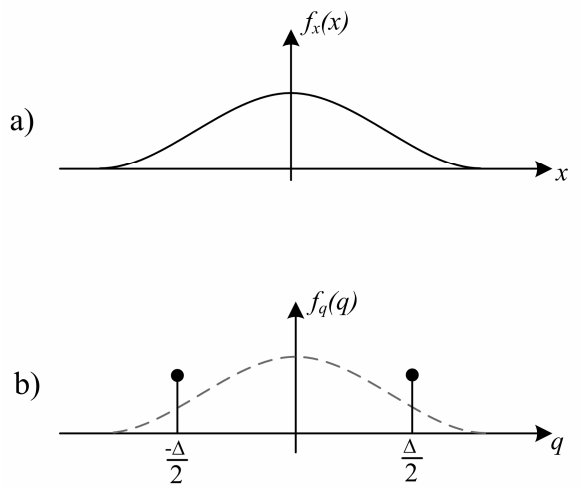


Figure 48: Input PDF (a) and output PDF (b), single-bit quantizer

For simplicity the following equations are normalized to the quantizer output, i.e. the output is ±1 and Δ=2. Since it is known that a DSM REQ with NTF zeros at DC forces the average output to equal the average input, the probability for the two output states can for any input level be described by the following two relations:

$$P(q=1)-P(q=-1)=x \ . \tag{81}$$

$$P(q=1)+P(q=-1)=1 \ . \tag{82}$$

Combining these two relations, the output PDF for any static input level is found to be:

$$f_q(q)=\delta(q-1)\frac{x+1}{2}+\delta(q+1)\frac{1-x}{2} \ . \tag{83}$$

The quantizer error is given by $e_q=q-x$ and its PDF is thus given by:

$$f_{e_q}(e_q)=\delta(e_q+x-1)\frac{x+1}{2}+\delta(e_q+x+1)\frac{1-x}{2} \ . \tag{84}$$

From the PDF each statistical moment of the error is easily found:

$$E\left[e_q\right]=0 \ . \tag{85}$$

$$E\left[e_q^2\right]=1-x^2 \ . \tag{86}$$

$$E\left[e_q^3\right]=-2x\left(1-x^2\right) \ . \tag{87}$$

It is noteworthy that as long as the average output equals the average input, this relation is constant *regardless of any applied dither*. This suggests that no dither changes the noise power modulation in a 1-bit DSM. It was supported by simulations in the first paper (Appendix 2). The *in-band* noise power modulation was found to be dominated by whether or not any idle-tones fell in-band, and by a power increase when the quantizer started to overload. This means that the sole purpose of dithering in a 1-bit DSM should be to eliminate tones. Since dither reduces the stable input range and increases the occurrences of overload, a binary DSM should *not* be dithered beyond rendering it sufficiently tone-free

3.5 Non-Overloading Delta-Sigma Modulators

As is now clear it is very difficult to ensure stable operation and control the quantization noise behaviour in a DSM REQ. To eliminate both tones and noise-power modulation completely the quantizer needs full TPDF dither. Being of width ±Δ this will eat up much of the input range in a few-bit quantizer, and if it can not be used while maintaining sufficient stable swing some non-ideal behaviour must be tolerated. Extensive simulations will be needed during design to ensure non-idealities don't deteriorate the output performance beyond what is acceptable.

If the quantizer is many-bit, there is on the other hand a simple method to *guarantee* stability for a strictly defined input range and – if desired – with full TPDF (or any) dither to eliminate idle-tones and noise-power modulation. This is achieved by designing the DSM using the non-overload method [123], which ensures no quantizer overload according to the range shown in ch.2.2. Repeating the DSM input-output relation in the *z*-domain,

$$Q(z) = STF(z)\cdot X(z) + NTF(z)\cdot E_q(z) \ , \tag{88}$$

the output of the quantizer *w*, is of course given by its output *q* minus its error e_q, meaning that it can be expressed as:

$$\begin{aligned} W(z) &= Y(z) - E_q(z) \\ &= STF(z)\cdot X(z) + \left(NTF(z) - 1\right)\cdot E_q(z) \ . \end{aligned} \tag{89}$$

Using the inverse *z*-transform, the corresponding time-domain expression is found:

$$w[n] = \sum_{k=0}^{\infty} stf[k]\cdot x[n-k] + \sum_{k=0}^{\infty} ntf[k]\cdot e_q[n-k] - e_q[n] \ . \tag{90}$$

The bounds for the peak amplitude of *w* can be found using the Cauchy-Schwartz inequality:

$$|w[n]| \le \left|\sum_{k=0}^{\infty} stf[k]\cdot x[n-k]\right| + \left|\sum_{k=0}^{\infty} ntf[k]\cdot e_q[n-k]\right| - |e_q[n]| \ . \tag{91}$$

$$\|w\|_\infty \le \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_q\|_\infty - \|e_q\|_\infty \ . \tag{92}$$

(92) is the more compact $\mathfrak{L}$-norm notation of (91). The $\mathfrak{L}$-norms of a vector are defined as:

$$\|x\|_p \overset{def}{=} \left(\sum_{n=0}^{\infty} |x[n]|^p\right)^{\frac{1}{p}} \ . \tag{93}$$

$$\|x\|_\infty \overset{def}{=} \max\left(|\mathbf{x}|\right) \ . \tag{94}$$

If the STF and NTF are FIR functions the peak quantizer input can be calculated exactly. If they are IIR functions they are infinitely long, but as long as they are BIBO-stable they converge to zero and the $\mathfrak{L}$-norms can be estimated with arbitrarily high precision using a large sample set. As long as the non-overload range *R* of the quantizer is larger than $||w||_\infty$ it never overloads. Consequently the non-overload requirement is:

$$|R| \geq \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_q\|_\infty - \|e_q\|_\infty \ . \tag{95}$$

If the quantizer has a dither input *v*, the dither sequence must be included in (90) and the procedure is easily repeated to find:

$$|R| \geq \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_q\|_\infty + \|ntf\|_1 \cdot \|v\|_\infty - \|e_q\|_\infty \ . \tag{96}$$

That this condition holds is a sufficient, but not necessary criterion for stability. It is a sufficient *and* necessary criterion to guarantee no overload.

Assuming a *B*-bit mid-thread quantizer like the one shown in fig.12 is used, the non-overload range is $|R| \leq 2^{B-1}-½$ normalized to the output. As long as there is no overload the peak error is limited to $||e_q||_\infty \leq ½$. If the NTF is basic mod*N*; $||ntf||_1 = 2^N$. The STF is usually unity and then if a stable input swing of half the quantizer input range – i.e. $||x||_\infty \leq 2^{B-2}$ – is desired, it's easy to calculate that the number of bits *B* must be at least:

$$2^{B-1} - \frac{1}{2} \geq 2^{B-2} + 2^N \cdot \frac{1}{2} - \frac{1}{2} \rightarrow B \geq N + 1 \ . \tag{97}$$

If the modulator has TPDF dither of width ±Δ, the requirement becomes much stricter:

$$2^{B-1} - \frac{1}{2} \geq 2^{B-2} + 2^N \cdot \frac{3}{2} - \frac{1}{2} \rightarrow B \geq N + 2 + \log_2\left(\frac{3}{2}\right) \ . \tag{98}$$

It is clear that since sufficiently high SQNR for hi-res audio will typically require mod3 or higher – see fig.20 – a non-overload DSM needs quite many bits in the REQ to get a good stable input range, especially if it is dithered. A conservative N^{th} order IIR NTF – e.g. one that is designed according to Lee's rule for 1-bit stability – will have a significantly smaller $\mathfrak{L}_1$-norm than mod*N*, perhaps reduced by 30-50%. This is not optimal with regards to SQNR vs. OSR, *but*; if both *N* and the OSR are sufficiently high for quantization noise to be negligible compared to other error sources, the non-overload modulator is very attractive. This is of course because it unlike other modulators can be made with guaranteed stability, no idle-tones and no noise power modulation. In the second paper (Appendix 4), non-overloading modulators are explored for different topologies and quantizer functions.

Chapter 4

Mismatch Shaping

As became clear in the previous chapter there are several advantages to using more than one bit in the DSM REQ. Apart from greater processing gain for low OSR, it is easier to ensure modulator stability, a larger input swing is tolerated and full RPDF or TPDF dither can be applied to eliminate tones and/or noise power modulation. The downside is that multi-bit DACs are not amplitude linear, which is the main reason why audio converters moved from LPCM to highly oversampled one-bit DSM in the first place. To achieve better than 10-12 bits resolution by physical matching alone is extremely difficult and dynamic element matching algorithms were introduced to spectrally shape the mismatch error contribution.

4.1 Mismatch Error Randomization

The term DEM in a data conversion context was introduced by Van De Plassche in 1976 [124] when he used redundant switching in a binary encoded DAC to improve its mismatch performance. In 1989 Carley showed an implementation of DEM in the sense it is known today, when he used a butterfly switching network to randomize the element selection in a thermometer encoded DAC, thus eliminating systematic INL [125]. This was an eight element DAC driven by a three-bit DSM REQ as shown in fig.49. In the figure thick lines indicate an ordinary multi-bit signal bus whereas tin lines are single (bit) lines.

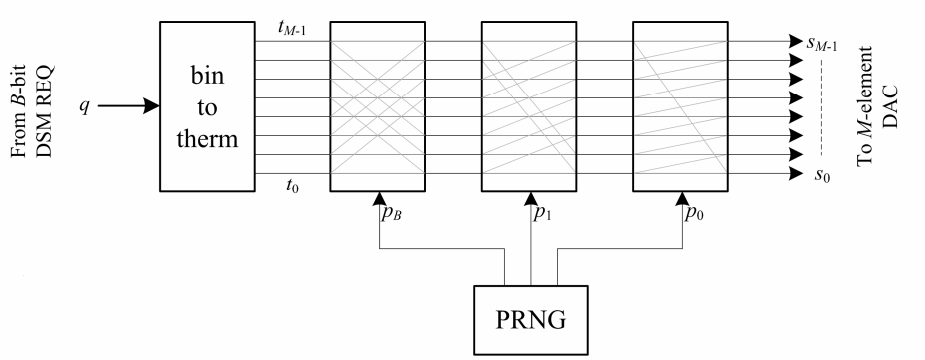


Figure 49: DAC element randomization, B=3 bit example

A B-bit DAC needs a B-bit switching network and PRNG to randomize the switching. As mentioned in ch.2 the number of elements M is typically 2^B or 2^B-1 depending on whether or not the REQ is symmetrical. For clarity the INL as a function of element weights is repeated:

$$INL(q) \overset{def}{=} \sum_{i=0}^{\hat{q}-1} w_i - \hat{q} \cdot \overline{w} \quad , \overline{w} = \frac{\sum_{i=0}^{M-1} w_i}{M-1} \; . \tag{99}$$

As before the INL is referred to the numeric quantizer output q and not given a unit. Input-referred – i.e. referred to x – its unit is Δ while the analog output y is referred to some reference current or voltage as explained in ch.2.

With randomization a random set of weights are assigned every time so one can't find an expression for the sample-to-sample error. But assuming the weights themselves are random variables with unity expectance value and variance σ_w^2, in other words that there are no graded or correlated errors[5], it is found that that the error expectance value $E\{e_w\}=0$ and the error variance as a function of q is:

$$\begin{aligned}\sigma_{e_w}^2(q) &= E\left\{\left(\sum_{i=0}^{\hat{q}-1} w_i - \hat{q}\cdot\bar{w}\right)^2\right\} \\ &= \hat{q}\cdot\left(1-\frac{\hat{q}}{M}\right)\cdot\sigma_w^2 \ .\end{aligned} \tag{100}$$

The maximum error variance occurs at the mid scale or q=0, when M/2 elements are switched on and the rest are turned off:

$$\sigma_{e_w}^2(0) = \frac{M\cdot\sigma_w^2}{4} \ . \tag{101}$$

This is the worst case error power with randomization. Since the elements are assumed to be Gaussian random variables, the error has a white spectrum. The Wiener-Khinchin theorem can be used to estimate the mismatch error PSD similarly to the quantization error:

$$S_{e_w}(\omega) = \frac{\sigma_{e_w}^2}{2\pi} \ . \tag{102}$$

This PSD can be integrated over the signal band to find the resulting SMNR.

4.2 Element Rotation Techniques

Element randomization efficiently turns mismatch non-linearity into a more benign white noise contribution, but it still does not facilitate very hi-res multi-bit conversion. Assume that the DAC is 4-bit with a mismatch standard deviation of 1% at the LSB-level (or σ_w=0.01); then the maximum SMNR with randomization is only around 60dB without oversampling or 80dB with an OSR of 128. This is clearly not sufficient for hi-res audio.

A few years after element randomization the concept of element *rotation* was introduced. This is based on the idea that since there is zero INL when all elements are in use, ensuring that every element contributes *equally over time* will cancel out the error. Several rotation algorithms were published in the early 90s, of which Individual Level Averaging [126] and in particular Data Weighted Averaging [127] turned out to be the most successful. More than a decade after its conception, DWA is still arguably the most popular DEM algorithm around. DWA element rotation is shown in fig.50. Note that the integrator *must* be a *B*-bit modulo integrator for the rotation to work as it should.

[5] This is a reasonable assumption if the DAC has good layout, utilizing common centroid techniques

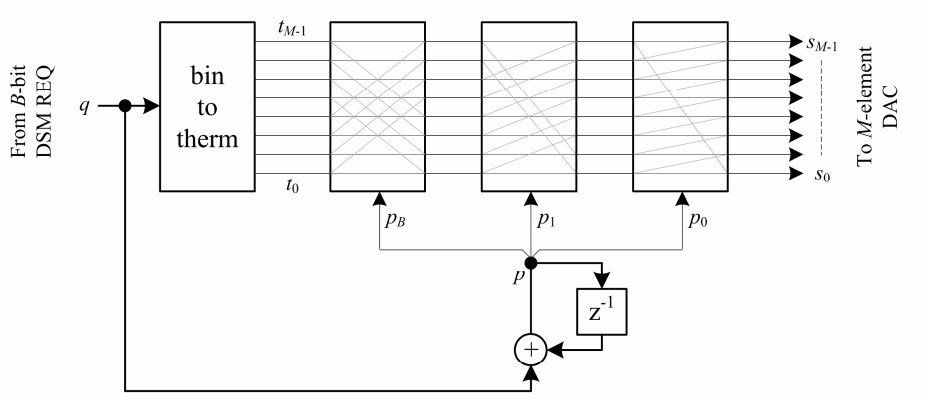


Figure 50: DWA DAC element rotation, B=3 bit example

Figure 51 exemplifies how element rotation uses each element equally over time. Over the course of five clock cycles it is seen that every element is used exactly twice, meaning that the net error cancels out over this time span. In real-life the averaging will of course mostly be slower, but as n grows every element will have contributed equally.

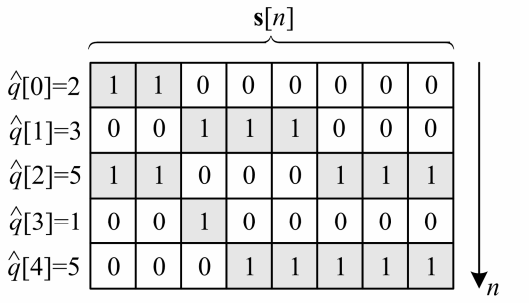


Figure 51: Element selection sequence with DWA

To simplify the description of the rotation scheme's mismatch shaping property, a vector notation was introduced in the fourth paper (Appendix 6). This notation defines an *element selection vector* **s** of length M, controlling the DAC element switching. The corresponding DAC error can then be written as:

$$e_w = \mathbf{s} \cdot \left(\mathbf{w} - \overline{w} \right) . \tag{103}$$

Here the vector **w** is the static element weight vector and the · operator is the vector dot product by conventional definition. To simplify the notation another vector **u** – also of length M – is defined as:

$$\mathbf{u}(a) \overset{def}{\rightarrow} u_i = \begin{cases} 1 & , 0 \le i \le a-1 \\ 0 & , a \le i \le M-1 \end{cases} . \tag{104}$$

If ordinary thermometer encoding is used the $\hat{q}$ lowest elements are always selected, meaning that the element selection vector can be described by:

$$\mathbf{s} = \mathbf{u}\left(\hat{q} \right) . \tag{105}$$

It follows from this that the DAC error for any given sample n is:

$$\begin{aligned} e_w[n] &= \mathbf{s}[n]\cdot(\mathbf{w}-\bar{w}) \\ &= \mathbf{u}(\hat{q}[n])\cdot(\mathbf{w}-\bar{w}) \\ &= \sum_{i=0}^{\hat{q}[n]} w_i - \hat{q}[n]\cdot\bar{w} \\ &= INL(q[n]) \ . \end{aligned} \tag{106}$$

This means any DAC INL translates directly to output distortion. In a DWA encoder on the other hand, the element selection is rotated by updating the starting point with a rotation pointer p (see fig.50), given by:

$$p[n] = (p[n-1] + \hat{q}[n]) \bmod M \ . \tag{107}$$

It is seen – use fig.51 for inspection if necessary – that the element selection vector can now be described as a function of the **u** vector as follows:

$$\mathbf{s}[n] = \begin{cases} \mathbf{u}(p[n]) - \mathbf{u}(p[n-1]) & , p[n] \geq p[n-1] \\ \mathbf{u}(M) + \mathbf{u}(p[n]) - \mathbf{u}(p[n-1]) & , p[n] < p[n-1] \end{cases} . \tag{108}$$

The vector $\mathbf{u}(M)$ indicates that the modulo pointer has wrapped around M. Using the same procedure as in (106) the DAC error is found to be:

$$e_w[n] = INL(p[n]) - INL(p[n-1]) \ . \tag{109}$$

Since $INL(M)=0$ this holds for both cases in (108). It means that the output distortion is a first order noise shaped function, since the z-transform gives:

$$E_w(z) = (1-z^{-1})\cdot INL(P(z)) \ . \tag{110}$$

Exact derivation of the error PSD requires exact knowledge of the statistics of the pointer. This is generally not trivial to obtain since the pointer is a modulo integral of the input. Approximations can however be made under certain conditions: If it is assumed that the input is a Gaussian-like random variable, $p[n]$ approximates a white random process. This makes it possible to use white noise estimation akin to Bennett's quantizer model. As long as the input signal is smaller than full-scale we also know that q is centred around and close to 0, so the worst case randomization estimate can be used as an approximation of the $INL(P(z))$ PSD. Under this assumption, the DWA DAC mismatch error PSD will be:

$$\begin{aligned} S_{e_w}(\omega) &= \frac{\sigma_{e_w}^2}{2\pi}\cdot\left|1-e^{\mathrm{i}\omega}\right|^2 \\ &= \frac{M\cdot\sigma_w^2}{8\pi}\cdot\left|1-e^{\mathrm{i}\omega}\right|^2 \ . \end{aligned} \tag{111}$$

With normalized input amplitude A, a B-bit REQ with the number of levels given by $M=2^B$ will have the corresponding signal-to-mismatch noise ratio:

$$SMNR = 10 \cdot \log_{10} \left(\frac{A^2 \cdot 2^B}{\frac{\sigma_w^2}{\pi} \int_{-\pi/L}^{\pi/L} \left| 1 - e^{i\omega} \right|^2 d\omega} \right) \text{ [dB] .} \tag{112}$$

As reviewed in ch.3.3, a DSM REQ designed according to Lee's rule typically has a max stable input of around -6dBFS or $A_{max}=0.5$, so $SMNR_{max}$ is easily found by insertion. Just like the DSM quantization noise estimate, this mismatch error estimate is based on the assumption that the input signal is a random process. In a real world scenario first order mismatch shaping has non-idealities quite similar to those of a first order DSM REQ. If the input is a DC-signal it is seen from (107) that $p[n]$ is a periodic function, meaning that a weighting error w_i will also appear periodically and create an error spectrum consisting of tones. Tones are not as severe as in a first order DSM REQ, since the input to the DEM block typically is a relatively few-level signal and contains a strong shaped quantization noise component. Spurs around the -100dB level are however to be expected and several techniques have been developed to alleviate tonality [128]-[130]. They typically dither of the rotation process, which reduces the tones at the cost of less efficient shaping. With very careful layout, DACs using first order mismatch shaping have achieved almost 18 ENOB [53],[56]. To improve further DEM must be evolved beyond first order shaping. A generalized analysis shows that second order DWA is theoretically trivial, but not easily to implement [131].

A simplified generalization requires for the *signal conservation rule* to be introduced. To preserve signal integrity the numeric output of the DEM encoder has to be equal to its input:

$$\sum_{i=0}^{M} s_i = \hat{q} \ . \tag{113}$$

This is obviously fulfilled with the DWA algorithm since $\mathbf{s}[n]=\mathbf{u}(p[n])-\mathbf{u}(p[n-1])$ and $p[n]-p[n-1]=q[n]$ for all n. In a second order extension of DWA – for convenience called 2DWA – it is desired that:

$$E_w(z) = \left(1 - z^{-1}\right)^2 \cdot INL\left(P(z)\right) \ . \tag{114}$$

From this the selection vector can be generally defined as:

$$\mathbf{s}[n] = c \cdot \mathbf{u}(M) + \mathbf{u}\left(p[n]\right) - 2 \cdot \mathbf{u}\left(p[n-1]\right) + \mathbf{u}\left(p[n-2]\right) \ . \tag{115}$$

The integer c is a carry variable saying how many times the pointer has wrapped around M. Pointer wrapping can now occur more than once, since to fulfil the signal conservation rule for (115) the pointer must be given by:

$$p[n] = \left(2 \cdot p[n-1] - p[n-2] + \hat{q}[n]\right) \bmod M \ . \tag{116}$$

The problem with direct implementation of this is that each element in the selection vector can now take other values than 0 or 1. For instance the same input sequence as in fig.51 will with 2DWA give the selection sequence shown in fig.52.

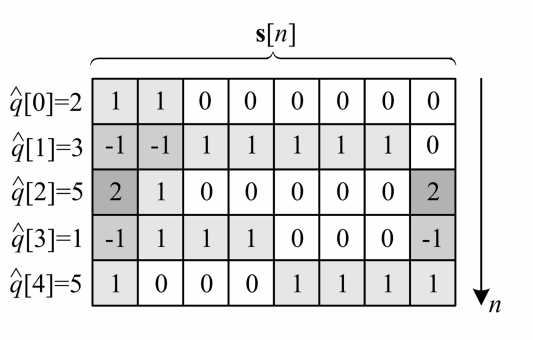


Figure 52: Element selection sequence with second order DWA

In this case each element needs to resolve four discrete levels. The '2' value can be obtained with a single element by running it at double the sampling rate, but it must still be ternary and is thus susceptible to internal mismatch. A solution to this problem was proposed and later patented as the Restricted 2DWA (R2DWA) algorithm [132]. In R2DWA an intermediate vector **it** is generated according to the second order noise shaping equation, and the algorithm then forces the selection vector **s** to take either '1' or '0' values, allocating ones to the $\hat{q}$ elements for which **it** has smallest entries.

for $n = 1 : \text{len}(data)$

$$\mathbf{it}[n] = 2 \cdot \mathbf{t}[n-1] - \mathbf{t}[n-2]$$

$$\mathbf{s}[n] = \text{all\_min}(\mathbf{it}[n], q[n])$$

$$\mathbf{t}[n] = \mathbf{it}[n] + \mathbf{s}[n]$$

end

In this pseudo code description the function **y**=all_min(**it**,*q*) allocates ones to *q* elements in **y** corresponding to those for which **it** has smallest values. In practice this introduces a compression in the 2DWA transfer function which must also be used in the feedback. The mismatch shaping function of R2DWA and other restricted second order DEM algorithms is thus on the general form:

$$H_{R2DWA}(z) = \frac{H_2(z)}{g(q) + (1 - g(q)) \cdot H_2(z)} \quad . \tag{117}$$

The function $H_2(z)$ denotes ideal 2DWA shaping or $H_2(z)=(1-z^{-1})^2$, and *g* is a less than unity compression factor. The value of *g* depends on the input signal and the number of levels in the DAC: If the input signal is small or the number of levels high, *g* is close to unity and the DEM efficiency is near ideal second order shaping. Simulations in [132] as well as paper 4 (Appendix 6) suggest a typical SMNR around 10dB worse than ideal 2DWA.

4.3 Other Techniques

Although DWA based rotation techniques were the breakthrough for DEM and consequently multi-bit DSM in hi-res applications, many publications have been made where the problem is attacked from a different point of view. This has led to some intriguing implementation approaches that are both more flexible and more hardware efficient than the rotation scheme in fig.50. In wide bandwidth applications it is desirable to use as low OSR as

possible and many bits in the REQ facilitates higher DSM processing gain for low OSR. Therefore the research activity in hardware efficient DEM techniques has been quite high.

An alternative way to understand the distortion generated by DAC mismatch is to view it as spectral leakage of single element switching sequences. The DSM generates an *M*-level signal, which in a thermometer encoder is divided into *M* two-level switching sequences routed to separate 1-bit DACs. An example of the element switching sequences s_0 to s_7 in an 8-level DSM DAC is shown in fig.53. Since the Fourier transform is linear, superposition gives that $\Sigma S_i(\omega)=Q(\omega)$. But if there is a weighting error in one of the elements the spectrum of its switching sequence will leak since:

$$\begin{aligned} E_{mis}(\omega) &= Q(\omega) - Y(\omega) \\ &= \sum_{i=0}^{M-1} (1-w_i)\cdot S_i(\omega) \,. \end{aligned} \tag{118}$$

Figure 53: Switching sequence for each element in a 3-bit DSM DAC

This means that the objective of DEM switching is to ensure that in addition to signal preservation or $\Sigma s_i=\hat{q}$, every switching sequence s_i *itself* has a shaped spectrum. Figure 54 shows the element switching sequences for the same input as fig.53 but now with DWA. A spectral analysis will reveal that every s_i has a spectrum consisting of a signal component and a first order shaped noise component. Thus DWA provides first order mismatch shaping.

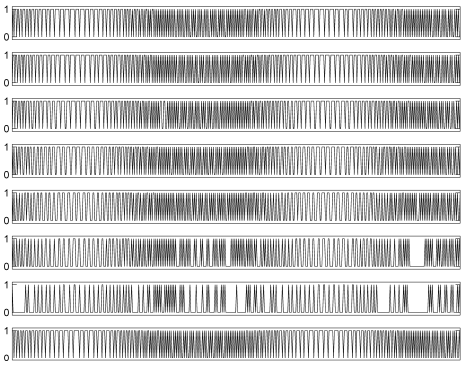


Figure 54: Switching sequence for each element in a 3-bit DSM DAC with DWA

The obvious question poised by this approach to mismatch distortion and DEM is consequently: How do you ensure that every switching sequence is spectrally shaped while keeping their combined sum equal to the input?

For a general two-element switching cell as shown in fig.55, where a control signal c determines whether the inputs are sent directly through the cell or if they are swapped, the outputs are necessarily given by:

$$s_1 + s_2 = a + b \ . \tag{119}$$

$$s_1 - s_2 = c \cdot (a - b) \ . \tag{120}$$

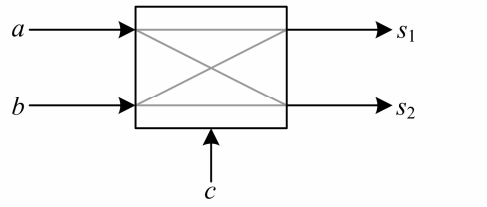


Figure 55: Two element swapper cell

Solving (119) and (120) for s_1 and s_2 it is found that they can be expressed as:

$$s_1 = \frac{1}{2}(a + b) + \frac{c}{2}(a - b) \ . \tag{121}$$

$$s_2 = \frac{1}{2}(a + b) - \frac{c}{2}(a - b) \ . \tag{122}$$

The property $s_1+s_2=a+b$ implies signal preservation. A weighting error in s_1 or s_2 introduces a non-unity signal gain, but more importantly leakage of $c\cdot(a-b)$ to the output. This means that as long as c is a shaped sequence the leakage is also shaped. To generate a first order shaped control sequence c can be done with simple logic as described in Adams' patent [133] used for Analog Devices converters. Higher order is much more complex, but the published R2DWA implementation is based on this type of swapper cells. Through induction it is found that ensuring *all* sequences are noise shaped requires for the swapper cells to be arranged in a complete swapping network like fig.56.

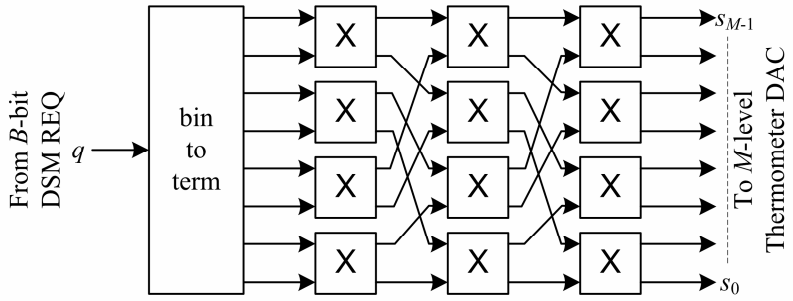


Figure 56: Swapping cell network for DEM, B=3

The "sum of shaped sequences" approach also led to an ingenious solution by Galton which significantly reduced the DEM complexity while maintaining high flexibility [134]-[136]. Both DWA and the swapper cell approach have $O(M\cdot\log_2 M)$ complexity for the switching network, in addition to a $O(M)$ complex thermometer encoder. Using a tree structure bit reduction logic, Galton reduced the thermometer encoder *and* DEM network to a single block just slightly above $O(M)$ complexity. This facilitates the use of more elements in the DAC.

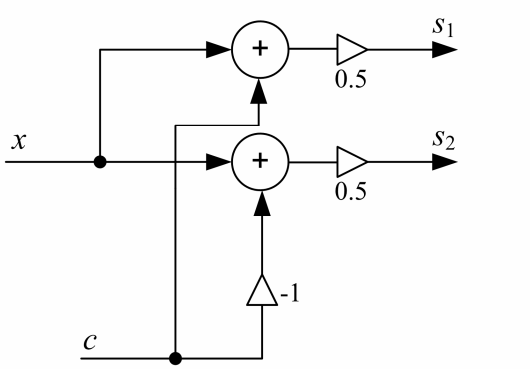


Figure 57: Data splitting and reduction for tree structure DEM

Imagine that an input signal x is split into two sequences as shown in fig.57, where we have:

$$s_1 = \frac{1}{2}(x+c)\ ,$$
$$s_2 = \frac{1}{2}(x-c)\ . \tag{123}$$

Since $s_1+s_2=x$ this structure is signal preserving. A weighting error between s_1 and s_2 means that c leaks to the output, so again c being shaped means the error is shaped. With a few other restrictions this structure can be used in a logic reduction tree. Firstly; to ensure both s_1 and s_2 are integers a restriction seen from (123) is that:

$$c = \begin{cases} \text{even if } x \text{ is even} \\ \text{odd if } x \text{ is odd} \end{cases}\ . \tag{124}$$

Furthermore; to enable bit reduction the outputs obviously have to be represented with less bits than the input, i.e. $s_{1,2} \leq 2^{B-1}$ for one bit reduction of a B-bit x. This is fulfilled as long as:

$$c \leq \min\left\{x, 2^B - x\right\}\ . \tag{125}$$

(125) is fulfilled for any positive B given $|c| \leq 1$. A control signal satisfying both (124) and (125) for every sample instant n can thus be made within the restriction:

$$c[n] = \begin{cases} 0 \text{ if } x[n] \text{ is even} \\ \pm 1 \text{ if } x[n] \text{ is odd} \end{cases}\ . \tag{126}$$

A simple modified 1-bit mod1 DSM can generate such a sequence that is also first order shaped, and can thus be used in a complete reduction tree with mismatch shaping. With a B-bit binary input and a set of 2^B two-level outputs where every output sequence s_i is first order shaped and $\Sigma s_i = q$, this structure with look like fig.58, showcasing a 3-bit example.

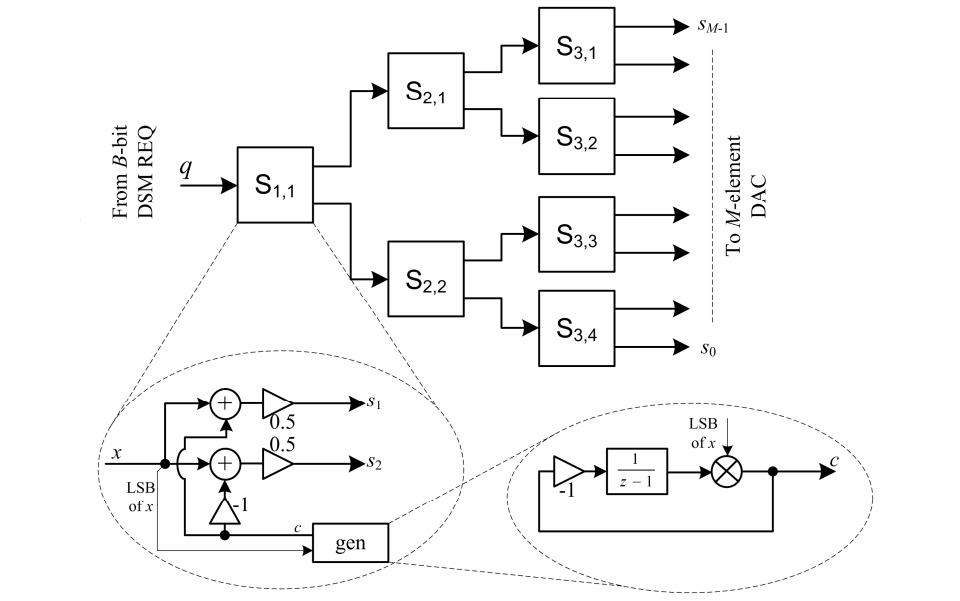


Figure 58: Complete reduction tree with first order mismatch shaping

Galton also showed simple logic for restricted second order shaping [135], but found higher order than this difficult to keep stable. A modified approach was recently shown [137] where higher order shaping is used for the first few switching layers (then the switching block input is more than two bits and the restrictions on c can be relaxed), while second order shaping is used for the last layers. Note that only the block that generates c has to be replaced to change the mismatch shaping function.

Most alternative (to DWA) algorithms are based on the sum of shaped sequences approach. One other that is noteworthy though not reviewed here, is the Schreier VQ-approach [138]

4.4 Segmented Mismatch Shaping

Even if DEM algorithms have become more efficient, a many-bit implementation will still be quite complex and chip area consuming. Especially for second order mismatch shaping this is true; at best a B-bit DAC will need 2^B modulators in the DEM encoder. The routing of a unit element DAC with many levels is also complex. An intuitive way to solve this would be to split the DAC into two sub-DACs with separate DEM encoders as shown in fig.59. Of the B bits B_0 LSBs are fed to the lower sub-DAC and the B-B_0 remaining MSBs are fed to the upper sub-DAC. The MSB DAC must then have an element weight of 2^{B_0} to give a correctly recombined output. The output is now sort of mid-way between thermometer code and binary code. A segmented DAC was first shown in 1979 [139], then without any DEM.

With segmentation there are now two smaller DEM blocks and less routing to implement. But although the DEM encoders linearize the sub-DACs, mismatch *between them* is not shaped. How this affects the output can be seen by making a signal flow diagram as shown in fig.60. For simplicity the sub-DACs are assumed linear (in reality they are DEM linearized), and inter sub-DAC mismatch is modelled as a weighting error $\alpha \neq 1$ in the LSB DAC.

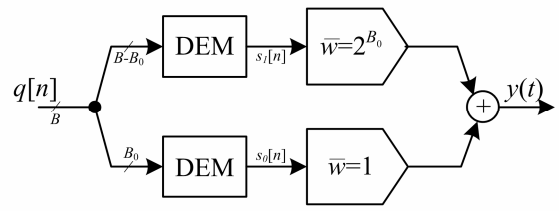


Figure 59: DEM and DAC segmentation

It is seen that splitting the data is equivalent to introducing a truncating quantizer and feed its truncated output to the MSB DAC. The truncation error is subtracted through the LSB DAC, meaning that this effectively acts as an error-compensation DAC. The MSB data is effectively right shifted by an amount equal to the number of bits shaved off – indicated by the 2^{-B0} gain element – which must be cancelled by a nominal MSB DAC gain of 2^{B0}. Ideally the compensation DAC has unity gain, but because of mismatch it is in reality some random variable α, making the compensation non-ideal[6]. The output is:

$$y = q + (1-\alpha)\cdot e \ . \tag{127}$$

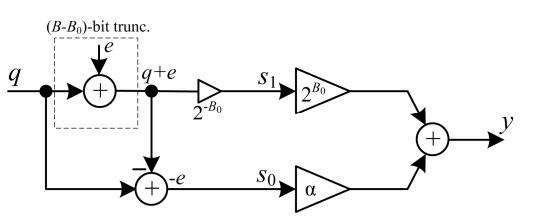


Figure 60: Equivalent signal flow diagram of segmented DAC

If the truncation is e.g. 4-bit and α=0.999, it means 0.1% of a 4-bit quantization error leaks to the output. A 4-bit quantization error suppressed by 60dB gives a total ENOB around 14. This is clearly insufficient for very hi-res applications, and a proposed solution was given by Adams in 1998 [140] where he replaced the truncation with a dedicated Segmentation-DSM (SDSM). The SDSM replaces the truncation and shapes the leaking error as shown in fig.61.

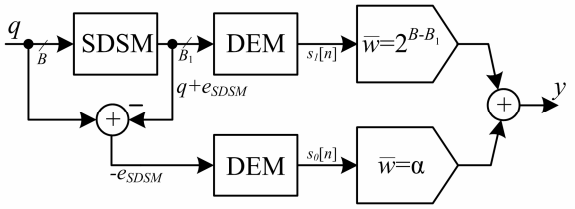


Figure 61: DEM and DAC segmentation with SDSM

The number of bits in the SDSM REQ is B_1. It thus scales the signal with a factor $2^{-(B-B1)}$ and the nominal weighting of the MSB DAC must compensate for this. The output is now:

$$y = q + (1-\alpha)\cdot e_{SDSM} \ . \tag{128}$$

Since e_{SDSM} is a shaped error the leakage caused by inter sub-DAC mismatch is also shaped. Conceptually this is very similar to the shaping of single element spectral leakage in a DEM-encoded unit-element DAC.

A disadvantage in replacing the truncation with an SDSM is that the peak error fed to the compensation-DAC grows in magnitude. With a truncator it is given that $B=B_0+B_1$ – as is evident from fig.59 – but when the REQ is situated inside a DSM the peak error increases. This means that the compensation-DAC needs more bits to accommodate a larger input swing. In his publication Adams used a first order error feedback SDSM with a (z-1) FIR NTF. Then $e_{SDSM}[n]=e[n]-e[n-1]$ and consequently $||e_{SDSM}||_\infty=2\cdot||e||_\infty$. This means that the compensation-DAC doubles in size.

[6] It is easily found that the variance of α is 2^{-B_0} times the nominal LSB-level DAC mismatch.

Generally for an $(z\text{-}1)^N$ FIR NTF the peak gain is 2^N, so in a unity STF SDSM it implies an N-bit increase of the compensation DAC. Generally $\|e_{SDSM}\|_\infty=\|ntf\|_1\cdot\|e\|_\infty$ and consequently – since $B_e=B\text{-}B_1$ – the compensation-DAC number of bits has to be at least:

$$B_0 \geq \left(B - B_1\right)\cdot\left\|ntf\right\|_1 \ . \tag{129}$$

This means that if B=8 and $(z\text{-}1)^2$ mismatch shaping is desired throughout the system, the most efficient DEM segmentation will be with a 5-bit SDSM, leading to both B_1 and B_0 being 5-bit signals.

Although not utilized in any published implementations known to the author, it is fully possible to use conservative non-overloading IIR NTFs in the SDSM. Since it has less peak gain such an NTF gives less additional cost from increasing the order. What improvements to expect with non-overloading IIR SDSMs compared to the FIR SDSMs previously used, is investigated in the third publication (Appendix 5). It reveals that the complexity penalty from increasing the SDSM shaping order can be significantly reduced.

Various structures for further DEM segmentation using SDSMs were investigated in the thesis by Steensgaard [141]. The most intuitive choice would be to just repeat the segmentation as is shown in fig.62. Steensgaard called this a "symmetrical tree structure" and he also explored "one-sided" and "asymmetrical" tree structures. Some are more efficient than others, but all create a DAC overhead that increases with the degree of segmentation.

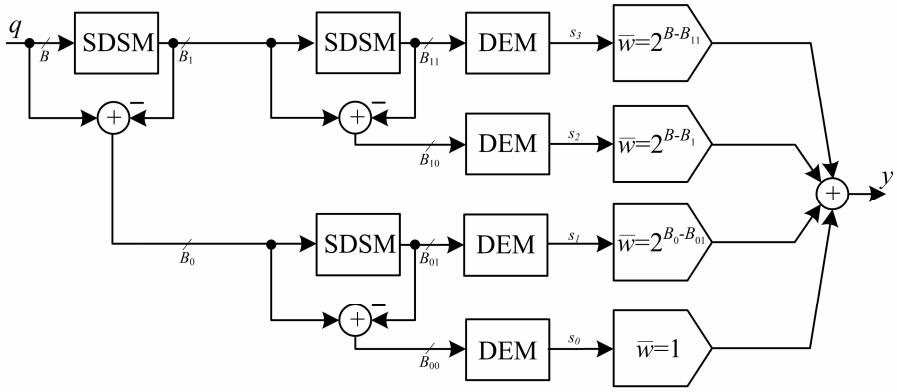


Figure 62: Two time DEM and DAC segmentation

On a final note a segmented version of the Galton tree structure has also been published [142]. In this the reduction tree is asymmetric and it thus generates something between a binary and thermometer type code. It is not reviewed here, but does have the same advantages as the regular Galton reduction tree, i.e. that it does not require a separate thermometer encoder and improves the hardware efficiency.

Chapter 5

Delta-Sigma and Dynamic DAC Errors

In chapter two the main categories of DAC errors; the quantization error, static errors, and dynamic errors, were reviewed. The next chapter explored how DSM REQ can facilitate few-bit or single-bit conversion with very low in-band quantization noise. It also showed benefits of using more than one bit, for instance that the in-band quantization noise can easily be made negligible while maintaining modulator stability. All multi-bit DACs have static non-linearity, but as the previous chapter reviewed DEM can be used to ensure very high resolution still.

This leaves the class of dynamic or waveform type errors. Chapter two showed the nature of such errors, but without relating them to the DSM REQ. In a DSM converter the DAC input is a coarsely quantized and noise shaped sample sequence, the nature of which significantly affects dynamic error sensitivity. Since it is generally not possible to analytically derive the DSM output sequence, it is neither possible to analytically derive dynamic errors. Simplified estimates can however be made and this chapter reviews the development of such.

5.1 Delta-sigma and Jitter Error Estimation

Jitter was introduced in chapter two as a waveform error caused by deviations in the sampling instant. It stems from the digital audio interface as well as noise and parasitics in the clock regeneration and distribution circuitry. The jitter pattern can appear signal correlated, as sinusoids, as white noise, and as pink noise. It was established that the jitter error can be approximated with an area error model as shown in fig.63, and that the error PSD then is:

$$S_{e_j}(\omega) \approx \frac{1}{T_s^2}\left[S_d(\omega) * S_j(\omega)\right] . \qquad (130)$$

Here *d* is the differentiated DAC input. It is desirable to have prediction models for the jitter distortion in a DSM DAC, so that qualified choices can be made for the DSM design. Development of such prediction models were featured in the fourth paper (Appendix 6).

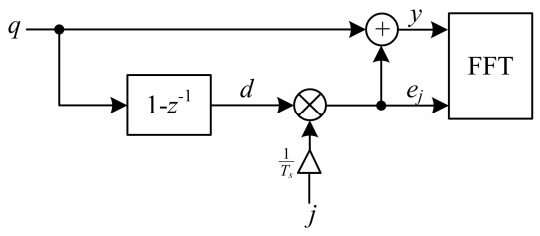


Figure 63: Area error model for jitter distortion analysis

To create a simple estimate of DSM DAC jitter distortion is most conveniently done through a frequency domain approach. As known the output sequence cannot be analytically derived, but we also know that in the frequency domain it approximates a spectrum consisting of a signal component and an independent shaped noise component. If the STF is unity in the signal band, the output PSD of the modulator can be expressed as:

$$S_q(\omega) \approx S_x(\omega) + \frac{\sigma_{e_q}^2}{2\pi} \cdot \left| NTF(\omega) \right|^2 \ . \tag{131}$$

The PSD of d or $S_d(\omega)$ is found through spectral differentiation:

$$S_d(\omega) \approx S_{dx}(\omega) + \frac{\sigma_{e_q}^2}{2\pi} \cdot \left| dNTF(\omega) \right|^2 \ , \ dNTF(\omega) \stackrel{def}{=} \left(1 - e^{-i\omega}\right) \cdot NTF(\omega) \ . \tag{132}$$

The jitter estimate makes use of the total power of d. This is found by integrating the PSD $S_d(\omega)$ across the whole frequency range $-\pi$ to π, or in other words find the spectral $\mathfrak{L}_2$-norm:

$$\sigma_d^2 \approx \sigma_{dx}^2 + \sigma_{e_q}^2 \cdot \left\| dNTF(\omega) \right\|_2^2 \ , \ \left\| H(\omega) \right\|_2 \stackrel{def}{=} \sqrt{\frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \left| H(\omega) \right|^2 d\omega} \ . \tag{133}$$

From the convolution theorem the power of e_j in (130) has to be:

$$\sigma_{e_j}^2 = \frac{1}{T_s^2} \cdot \sigma_d^2 \cdot \sigma_j^2 \ . \tag{134}$$

The cases of white random jitter and sinusoid sideband jitter were explicitly considered in the paper since these are most likely to cause audible distortion or noise[7]. If the jitter PSD is white, the jitter error PSD will also be white since it stems from convolution. This means that $1/L$ of the total error power (134) fall in-band, and the in-band jitter noise power is:

$$\hat{\sigma}_{e_j}^2 = \frac{1}{L \cdot T_s^2} \cdot \left(\sigma_{dx}^2 + \sigma_{e_q}^2 \cdot \left\| dNTF(\omega) \right\|_2^2 \right) \cdot \sigma_j^2 \ . \tag{135}$$

If the signal component is a sinewave with output normalized peak-to-peak amplitude $A \cdot 2^B$, and $\omega_x << \pi$ so that $A_{dx} \approx A_x \cdot \omega_x$, the in-band SJNR will be:

$$SJNR \approx 10 \cdot \log_{10} \left(\frac{\dfrac{A^2 \cdot 2^{2B}}{f_{s\_in}^2 \cdot L}}{\left(A^2 \cdot 2^{2B} \cdot \omega_x^2 + \dfrac{2}{3} \left\| dNTF(\omega) \right\|_2^2 \right) \sigma_j^2} \right) \ . \tag{136}$$

If the number of bits B is very large, e.g. in a hi-res LPCM converter[8], the denominator is dominated by the signal term and jitter noise approximates that of ordinary sampling jitter [63]. With few bits the quantization error term dominates the denominator in (136), and the SJNR reduces by 6dB for each bit removed. Achieving hi-res performance is very difficult with few bits since the phase noise variance must then be extremely low. Figure 64 illustrates this by showing $SJNR_{\max}$ for varying numbers of levels assuming a peak input of -6dbFS ($A = A_{\max} = 0.5$). The NTF of the DSM REQ is also the same in all examples and designed according to Lee's Rule. The input sampling frequency f_{s_in} is 44.1kHz and the jitter $50ps_{RMS}$.

[7] Pink jitter noise is likely to be masked, and in this context in-band jitter sidebands behave the same whether they are correlated or uncorrelated.

[8] A non-modulating REQ will have an NTF of 1.

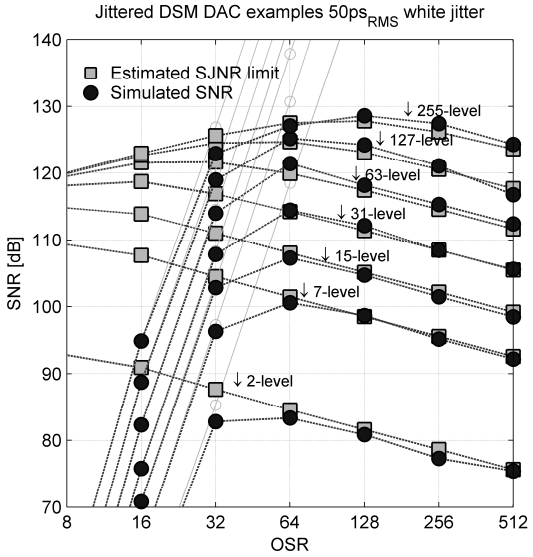


Figure 64: $SJNR_{max}$ example, 50ps white jitter

The figure compares $SJNR_{max}$ according to the estimate with simulated SNR in a high order DSM DAC. For low OSR the quantization noise dominates the simulated error while for high OSR the performance is jitter limited. It is seen that with $50ps_{RMS}$ white jitter the DAC needs quite many bits to maintain hi-res audio performance.

Sinusoid jitter leads to sideband distortion since convoluting the power spectral densities means discrete jitter components are multiplied with the spectral components in d, which include a signal component and shaped noise. If the signal x is sinusoid, straightforward multiplication through the angle sum and difference identities gives resulting modulation products at $\omega_x \pm \omega_j$ with amplitude:

$$A_{e_j}^{(\omega_x \pm \omega_j)} = \frac{A_d \cdot A_j}{2 \cdot T_s} \approx \frac{A_x \cdot \omega_x \cdot A_j}{2 \cdot T_s} \quad . \tag{137}$$

No component in the quantization noise contains enough power by itself to create discernible modulation products with sinusoid jitter, so the total distortion approximates (137) and is equivalent to sampling jitter. Since convolution is linear, calculation of jitter noise and jitter sidebands from a composite jitter spectrum can be done separately before adding them together. Figure 65 shows simulated output spectra of a DSM DAC with sinusoid, discrete, and mixed jitter. It is seen that combining them does not affect the contribution of each.

In conclusion jitter sideband distortion is not affected by the DSM, but to maintain high SNR in the presence of white PSD phase noise it must be many-bit or out-of-band noise must be removed while in discrete time. A switch-cap filtering DAC does the latter; and the differentiated PSD at the discrete-to-continuous interface – or SCF output – will be:

$$S_d(\omega) \approx S_{dx}(\omega) + \frac{\sigma_{e_q}^2}{2\pi} \cdot \left| d\left[NTF(\omega) \cdot H_{SCF}(\omega) \right] \right|^2 \quad . \tag{138}$$

$H_{SCF}(\omega)$ is the low-pass response of the switch-cap filter. The advantage of SC-filtering was assessed experimentally by Fujimori [53], but can also be estimated with the above method. As an alternative it is possible to explore other types of reconstruction than zero order hold. Hawksford suggested raised-cosine reconstruction [143], but due to the difficulty of a hi-res implementation it has not been seen in commercial applications to my knowledge.

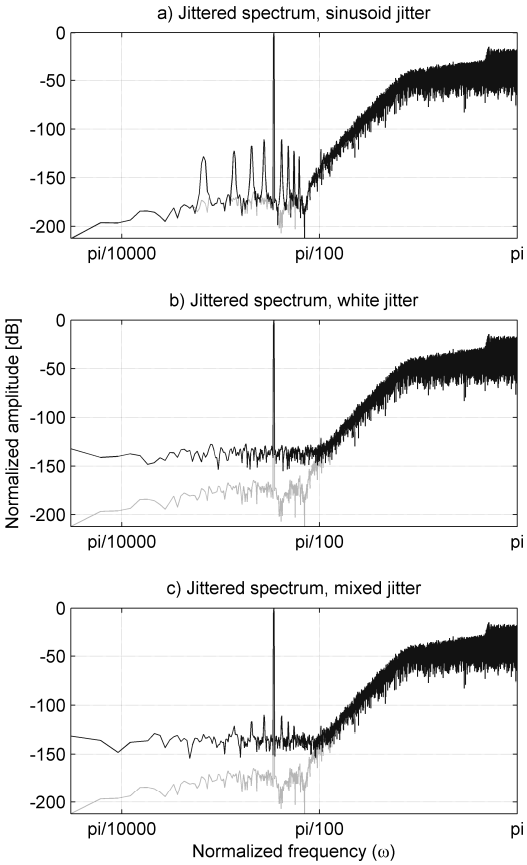


Figure 65: Jittered spectrum with a) sinusoid, b) white and c) mixed jitter

5.2 Delta-sigma and Switching Error Estimation

In chapter 2.7 it was established that the error waveform or ISI generated by switching errors e_{on} and e_{off} in a DAC, can be approximated using an area error model:

$$e_{ISI}[n] \approx \begin{cases} d_n \cdot e_{on}, & d_n \geq 0 \\ d_n \cdot e_{off}, & d_n < 0 \end{cases}. \tag{139}$$

It is seen that this error is proportional to d and scales with either e_{on} or e_{off} depending on whether d is positive or negative. Thus $e_{on}=e_{off}$ means that e_{ISI} is a linear product of d and the ISI is benign. On the other hand, if $e_{on}\neq e_{off}$ the error waveform is asymmetrical around $d=0$ and in other words constitutes a non-linearity.

Just like the jitter error approximation, the ISI error approximation is based on superposition of spectral components in the differentiated DSM output, as given by the additive noise model. Assessing first the signal component: The assumption $x_n=A\cdot\cos(\omega_x n)$ and $\omega_x<<\pi$ gives that $d_n\approx -A\cdot\omega_x\cdot\sin(\omega_x n)$, and (139) can thus be rewritten to:

$$e_{ISI}[n] \approx -\begin{cases} A\cdot\omega_x\cdot\sin(\omega_x n)\cdot e_{off}, & 0 \leq \omega_x n < \pi \\ A\cdot\omega_x\cdot\sin(\omega_x n)\cdot e_{on}, & \pi \leq \omega_x n < 2\pi \end{cases}. \tag{140}$$

Since (140) is infinite and periodic in $\omega_x n=2\pi$ its Fourier series can be developed, which was showcased in analysis by Clara et al. [79] and results in even harmonic spectral components with amplitude:

$$A_{ISI}^{(k\cdot\omega_x)} \approx \begin{cases} \dfrac{2\cdot\left|e_{off}-e_{on}\right|}{\pi(k+1)(k-1)}\cdot A\cdot\omega_x & ,\ k=2,4,6... \\ 0 & ,\ \text{otherwise} \end{cases} . \tag{141}$$

Paper four extended this analysis to also assess the impact of the shaped quantization noise component e_{dsm}. As is known either e_{off} or e_{on} multiplies with d depending on whether or not its instantaneous value is above or below zero. A time sequence of e_{dsm} can again not be found analytically, but finding its sign means subjecting it to 1-bit unshaped quantization. 1-bit unshaped quantization of a shaped noise sequence effectively renders it white, and whether e_{off} or e_{on} multiplies with d is thus something given by a random process with a white PSD. Since the NTF has zeros at DC this process can be assumed zero-mean and its total power is approximately $|e_{off}-e_{on}|^2$. The expression for the approximate total power of d is already known – see (133) – and the total ISI error power is follows from it:

$$\sigma_{ISI}^2 = \sigma_d^2\cdot\left(e_{off}-e_{on}\right)^2 \approx \frac{1}{12}\left\|dNTF\right\|_2^2\cdot\left(e_{off}-e_{on}\right)^2 . \tag{142}$$

Since the sign sequence is approximately white, it means the ISI error it produces is also approximately white. In-band noise power is therefore $1/L$ of the total noise power, and in-band SSNR *disregarding distortion from the signal component* is thus approximately:

$$SSNR \approx 10\cdot\log_{10}\left(\frac{\dfrac{A^2\cdot 2^{2B}}{f_{s\_in}^2\cdot L}}{\dfrac{2}{3}\left\|dNTF\left(\omega\right)\right\|_2^2\cdot\left(e_{off}-e_{on}\right)^2}\right) . \tag{143}$$

With a constant element on-error $-e_{on}$ and off-error e_{off}, the SSNR increases by 6dB per bit since the switching activity caused by the differentiated DSM noise remains constant while the maximum signal swing increases proportionally.

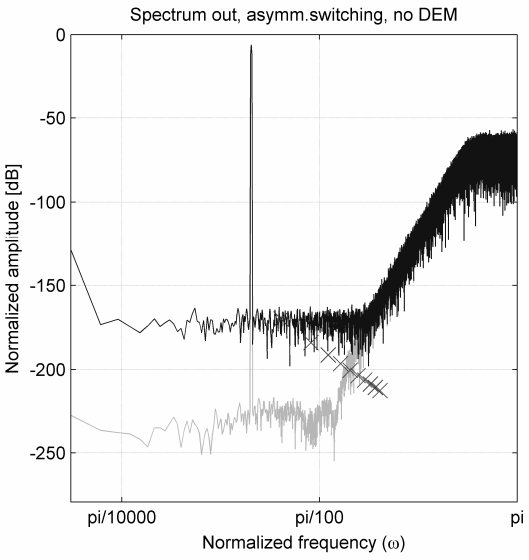


Figure 66: Simulated spectrum, 10ps switching asymmetry

Figure 66 shows a simulated output spectrum from a DSM DAC with asymmetric switching. The switching error area in this simulation is calculated from linear slewing and a rise-time and fall-time asymmetry of 10ps. The grey trace is the ideal DSM output and the black trace is the DAC output. Harmonic signal distortion components estimated from (141) are shown as markings, and as seen all are buried in the noise floor. Switching asymmetry can thus be approximated as a white error with an error power estimated from (142).

Comparisons of this estimate with performance simulations give the result of fig.67. The bottom to top trace shows 7-level to 255-level DACs with the same relative element switching error; that is linear slewing with 10ps rise-time and fall-time asymmetry in every element. For low OSR the quantization noise dominates, while for high OSR the ISI limits performance. Results are now shown for the 2-level DAC since the ISI models were made to facilitate DEM and thus had to be multi-level, but can obviously be expected to be worse.

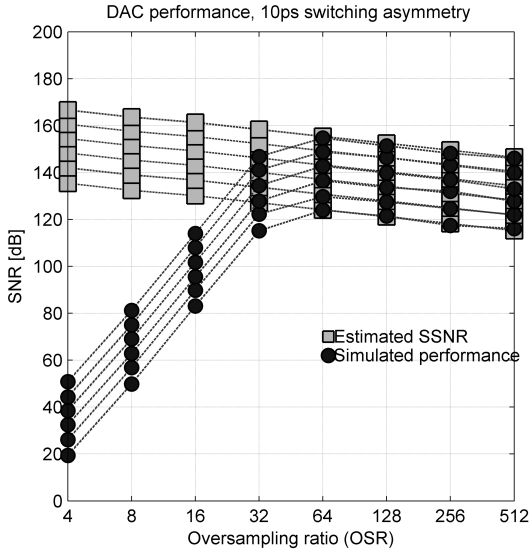


Figure 67: Simulated $SSNR_{max}$ example, 10ps switching asymmetry

As seen the simple estimate matches the simulated performance very well when the latter is limited by switching asymmetry. It should be noted though that like the others this estimate is based upon simplified approximations. Notably the additive noise model is used, but also the sign sequence – and from it the error sequence – is assumed to have a white PSD.

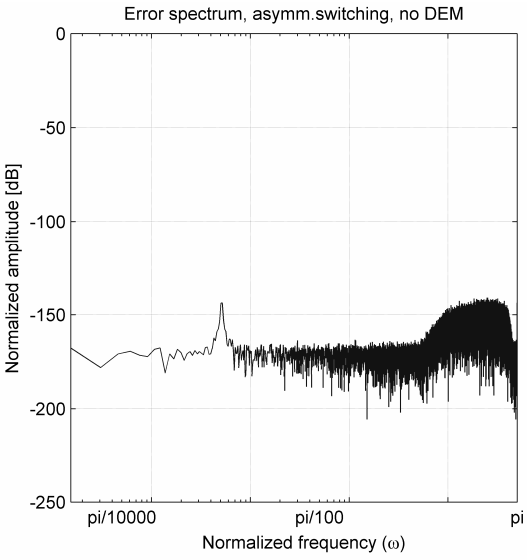


Figure 68: Simulated ISI error spectrum

Figure 68 shows the extracted error spectrum from the simulation used in fig.66. As can be seen the error is in reality not entirely white but contains some residuals of the signal and out-of-band noise components. Nonetheless the estimate gives a good prediction of the SSNR. Note also that just like for jitter distortion it will be advantageous to use a many-bit system.

With DEM the switching activity is quite different, which is clearly seen from fig.51 as well as fig.54. Assuming DWA is used, (108) and (139) gives the following relation between the DSM output time sequence and the switching error:

$$e_{ISI}[n]=\begin{cases}\hat{q}[n]\cdot e_{on}-\hat{q}[n-1]\cdot e_{off} & ,\ \hat{q}[n]+\hat{q}[n-1]\le M\\ \left(M-\hat{q}[n-1]\right)\cdot e_{on}-\left(M-\hat{q}[n]\right)\cdot e_{off} & ,\ \hat{q}[n]+\hat{q}[n-1]>M\end{cases}. \quad (144)$$

Evaluating first the signal component – i.e. assuming $q[n]=A\cdot\sin(\omega_x n)$ where $\omega_x n<<\pi$ – it is found that the ISI error will be:

$$e_{ISI}[n]=\begin{cases}\left(\frac{M}{2}+A\sin\left(\omega_x n\right)\right)\cdot e_{on}-\left(\frac{M}{2}+A\sin\left(\omega_x\left(n-1\right)\right)\right)\cdot e_{off}\ ,\ 0\le\omega_x n<\pi\\ \left(\frac{M}{2}-A\sin\left(\omega_x\left(n-1\right)\right)\right)\cdot e_{on}-\left(\frac{M}{2}-A\sin\left(\omega_x n\right)\right)\cdot e_{off},\ \pi\le\omega_x n<2\pi\end{cases}. \quad (145)$$

Just like without DEM, $e_{on}=e_{off}$ means that e_{ISI} is a linear function of the signal. Fourier series development of (145) – also shown in [79] – results in even harmonics with amplitude:

$$A_{ISI}^{(k\cdot\omega_x)}\approx\begin{cases}\frac{2\cdot\left|e_{off}-e_{on}\right|}{\pi(k+1)(k-1)}\cdot A & ,\ k=2,4,6...\\ 0 & ,\ \text{otherwise}\end{cases}. \quad (146)$$

To develop a spectral estimate for the additional in-band noise that is caused by asymmetric switching of e_{dsm}, would be difficult since (144) spectrally constitutes a non-linear filter. But as simulation shows in fig.69; harmonic distortion is very dominant, and is clearly the limiting performance factor in the sense that asymmetric switching makes the SFDR unacceptable long before the SNR. Estimation of additional in-band noise was consequently not pursued.

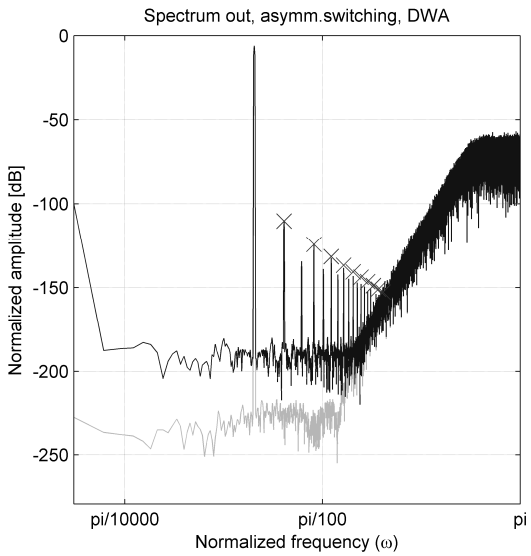


Figure 69: Simulated spectrum, 10ps switching asymmetry, DWA

It is seen that although the estimate (146) only predicts even harmonics there are also some weaker albeit clearly visible *odd* harmonics, which are not predicted by the signal analysis nor mentioned in [79]. Remember that the estimate (146) only assesses the signal component and does not take into account that the DAC input is generated by a DSM. The odd harmonics are probably caused by e_{dsm} in reality *not* being independent of the input, although the additive noise model assumes it is. So in addition to causing switching noise the DSM error will also make asymmetric switching cause some odd harmonic content.

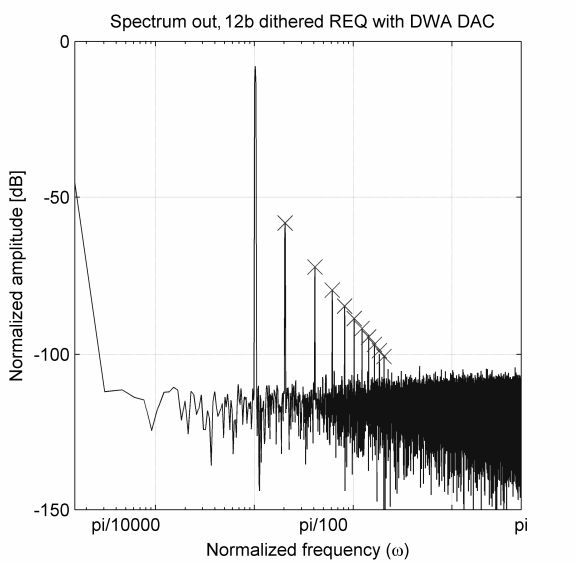


Figure 70: Simulated spectrum of LPCM DAC with DWA

Figure 70 shows the output spectrum of a DWA DAC with asymmetric switching, but instead of a DSM REQ the DAC input is now generated by a TPDF dithered 12-bit LPCM REQ. The switching asymmetry is significantly increased to make the distortion clearly visible in the spectrum. As seen odd harmonics are now *not* present and simulations give a distortion matching the estimate in (146) and [79]. Although a DSM REQ causes the distortion to also contain some odd harmonics and additional in-band noise; the ISI error with DEM is still dominated by even harmonics and (146) will accurately predict the SFDR.

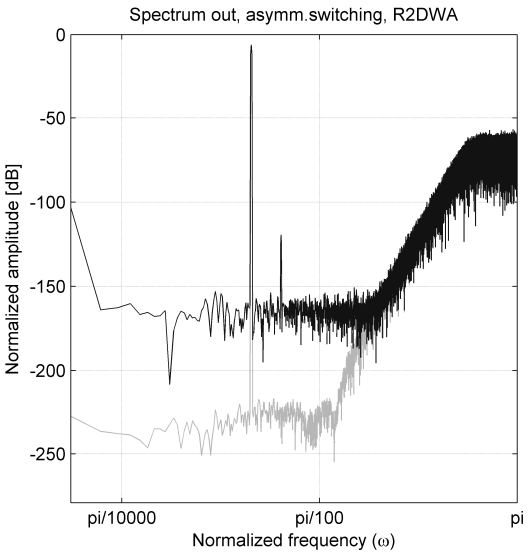


Figure 71: Simulated spectrum, 10ps switching asymmetry, R2DWA

With higher order DEM it is not possible to derive the switching sequence and it would therefore be extremely difficult to create good estimates for switching asymmetry distortion. Simulations do show that the ISI error will be dominated by in-band switching noise and a strong second-harmonic component as seen in fig.71. The SFDR is approximately 10dB better than with first order DEM, and the harmonic spectrum is more benign. Thus second order DEM will be superior over first order also to reduce ISI distortion.

5.3 Techniques for Reducing Dynamic Errors

Back in the 1980s when 1-bit delta-sigma was the dominating design paradigm for high resolution audio converters, it was quickly acknowledged that switching errors would be a limiting factor for the performance [49]. Investigation into techniques to reduce this problem followed shortly thereafter.

5.3.1 Return to zero

A solution that was soon proposed for this was the same that is often used to eliminate ISI in digital transmission channels, namely return-to-zero switching. An RZ DAC simply resets every element within each sample, creating an output waveform as shown in fig.72.

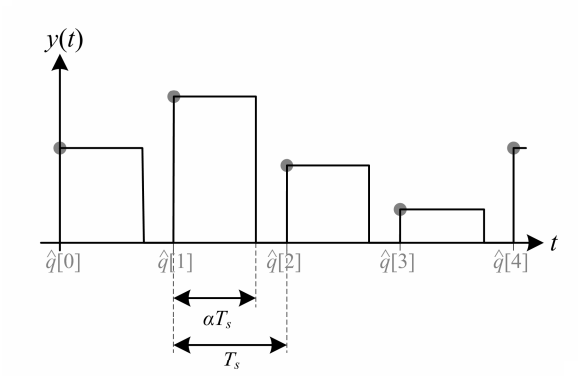


Figure 72: Return-to-zero waveform

The elements are switched on for a given fraction of the sample period α<1. Now, regardless of the value of the input sample, a number of elements equal to this value are turned both on and off within one sample period. This means the error expression reduces to:

$$e_{ISI}[n] = \hat{q}[n] \cdot \left(e_{off} - e_{on} \right). \tag{147}$$

Now the ISI error is a linear function of the input also if switching is asymmetric, meaning that it is benign. Thus ISI is eliminated fully, as long as the settling is complete. But even though ISI is eliminated, RZ switching does have some major disadvantages.

Since the sample period must be divided into a reconstruction phase and a reset phase, internal clock speeds must be higher than the sampling rate. There are also high frequency components produced at the output, which may fold down due to non-linearities and insufficient filtering. Switching losses are increased and the output power reduced by a factor α^2 for a given element current. But most importantly; since the output resets to zero for each sample, the RZ DAC is highly sensitive to random clock jitter. Assuming instantaneous jitter values at nT and $(n+\alpha)T$ are uncorrelated random values, the jitter error PSD is approximately:

$$S_{e_j}(\omega) \approx \frac{2}{T_s^2}\left[S_{\hat{q}}(\omega) * S_j(\omega)\right] . \tag{148}$$

We know that $\hat{q}[n] = M/2 + Q(x[n])$ and thus its DTFT is:

$$\hat{Q}(\omega) \approx \frac{M}{2}\sum_{k=-\infty}^{\infty} 2\pi \cdot \delta(\omega + 2\pi k) + X(\omega) + E_{dsm}(\omega) . \tag{149}$$

If the jitter is white we have from (148) that the error is also white, with in-band power:

$$\sigma_{e_j}^2 = \frac{2}{L \cdot T_s^2} \cdot \sigma_{\hat{q}}^2 \cdot \sigma_j^2 , \tag{150}$$

and the integrated PSD or spectral $\mathfrak{L}_2$-norm of $\hat{q}$ is found to be:

$$\begin{aligned} \sigma_{\hat{q}}^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{Q}(\omega)\, d\omega \approx \frac{M^2}{4} + \sigma_x^2 + \sigma_{e_q}^2 \cdot \left\| NTF(\omega) \right\|_2^2 \\ \rightarrow \sigma_{e_j}^2 &\approx \frac{2}{L \cdot T_s^2} \cdot \left(\frac{M^2}{4} + \sigma_x^2 + \sigma_{e_q}^2 \cdot \left\| dNTF(\omega) \right\|_2^2 \right) \cdot \sigma_j^2 \end{aligned} . \tag{151}$$

This results in an SJNR for sinusoidal input signals that is given by:

$$SJNR_{RZ} = 10 \cdot \log_{10}\left(\frac{\dfrac{\alpha^2 \cdot A^2}{2 \cdot f_{s\_in}^2 \cdot L}}{\left(2 + A^2 + \dfrac{2}{3 \cdot 2^{2B}}\left\| NTF \right\|_2^2\right)\sigma_j^2} \right) . \tag{152}$$

Figure 73 shows SJNR estimates and SNR simulations of a RZ DAC with duty-cycle α=0.8. It is seen that for a 2-level DAC the sensitivity to random jitter roughly doubles since there are two jittered edges instead of one and signal power is reduced by α^2. With many levels the waveform is always reset from mid-scale to zero, which dominates the jitter error area and means that the first term dominates the denominator in (152). The SJNR will then not increase as the number of levels is increased. This implies that RZ switching makes the use of many-bit DACs pointless, which is confirmed by the simulated jitter performance.

RZ switching might on the other hand be advantageous for low frequency sinusoidal jitter, since the instantaneous jitter values $j(nT)$ and $j((n+\alpha)T)$ are then very similar in amplitude. This means that the area error from switching on is nearly cancelled by the area error from switching off. An approximation for sideband distortion with RZ can be derived identically to the NRZ case and is found to be:

$$A_{e_j}^{(\omega_x \pm \omega_j)} = \frac{A_x \cdot A_{dj}}{2 \cdot T_s} \approx \frac{A_x \cdot A_j \cdot \omega_j \cdot \alpha}{2 \cdot T_s} . \tag{153}$$

There is also a distortion component at ω_j due to mixing with the offset. Since the offset is $M/2$ its amplitude will be as given in (154).

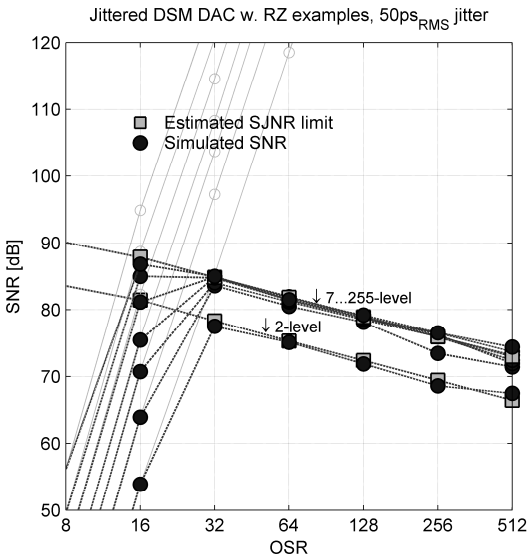


Figure 73: $SJNR_{max,}$ 50ps white jitter and RZ DAC

$$A_{e_j}^{(\omega_j)} \approx \frac{M \cdot A_j \cdot \omega_j \cdot \alpha}{T_s} \ . \tag{154}$$

Actually the susceptibility to low frequency sinusoidal jitter is somewhat improved with RZ switching compared to NRZ, but since the sensitivity to white or wide-band jitter is so high RZ is only really usable for 1-bit DSM DACs.

5.3.2 Dual return-to-zero and time interleaving

Since traditional RZ switching ruins the gain in jitter sensitivity from using multi-bit DSM conversion, developers and researchers quickly ventured into research on methods for ISI-elimination that preserve the output waveform. Adams proposed a variation called dual-RZ, which he introduced with the same innovative DAC design that also introduced segmented DEM [140]. Dual-RZ was described closer in a subsequent JSSC publication [144]. The design uses two RZ sub-DACs clocked in opposite phase and sums their outputs to form a replica of the input waveform as shown in fig.74.

If each sub-DAC element is associated with a turn-on error $-e_{on}$ and a turn-off error e_{off}, the combined error from the two RZ sub-DACs becomes:

$$e_{ISI}[n] = 2 \cdot \hat{q}[n] \cdot \left(e_{off} - e_{on}\right) \ . \tag{155}$$

This means that ISI is eliminated as long as settling is complete, and the two sub-DACs are driven by *the same* clock signal. If they are, any deviations in the clock transition will affect both identically and the reproduced waveform is an input waveform replica as shown in the figure regardless of jitter. The jitter sensitivity is thus the same as for an ordinary NRZ DAC. Disadvantages with this scheme include it requiring two RZ DACs meaning it has double the complexity and even higher switching losses. Additionally, synchronization of the two sub-DACs will be very critical to the final reproduced waveform and the converter's performance.

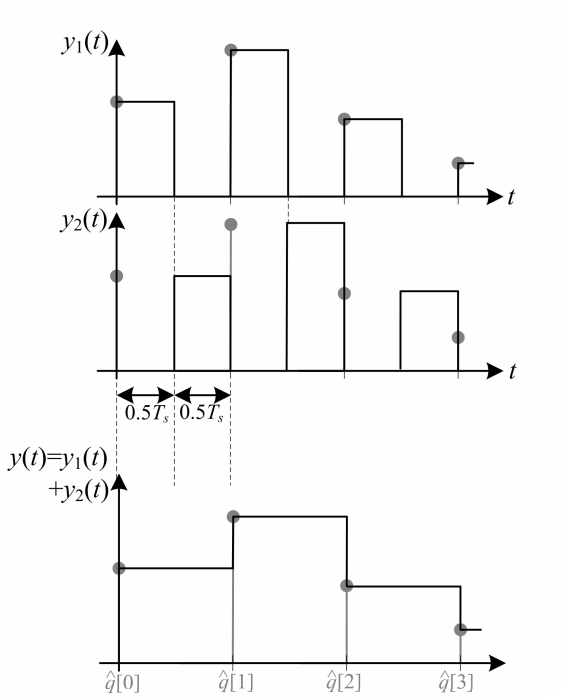


Figure 74: Dual-RZ waveform

Another approach, that was first proposed by Steensgaard [141] and in a variation used for a more recent high-speed DSM DAC design [145], is DAC time-interleaving. Straightforward sample-interleaving cannot be used since mismatch between sub-DACs then produces output distortion. But this can be dealt with by modifying the DEM scheme [145] or by interleaving in such a way that both sub-DACs contribute equally to every sample, shown in fig.75 [141].

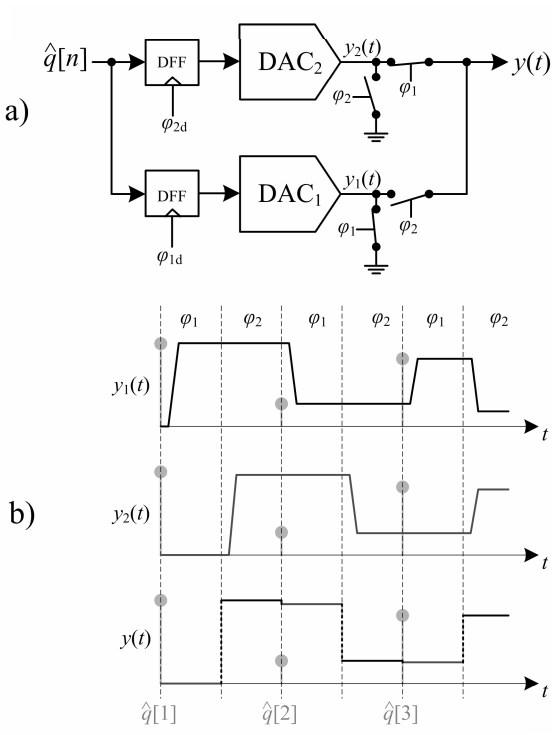


Figure 75: DAC time-interleaving, a) functional diagram, b) waveform

In this approach the sub-DACs are not RZ, but they are allowed to settle *before* they are connected to the output. This means that the sub-DACs can be slow and their dynamic behaviour sluggish without it affecting the output waveform. The dynamic behaviour of the output switches *will* on the other hand affect the waveform and may cause ISI distortion. Its transitions are shown as dotted lines in the figure. The design is however much improved over regular NRZ since it is much easier to control the switching behaviour of a single output switch than a score of DAC elements. An implementation suggestion is featured in [141].

5.3.3 Semidigital filtering DAC

Back when the norm was to use 1-bit DSM REQs, several ways to improve the jitter performance were explored and one of the more useful proposals was the semidigital filtering DAC [146]. By arranging several DAC elements as coefficients in a semidigital FIR filter, a multi-level output signal could be created where mismatch did not affect the DAC linearity. This concept is shown in fig.76.

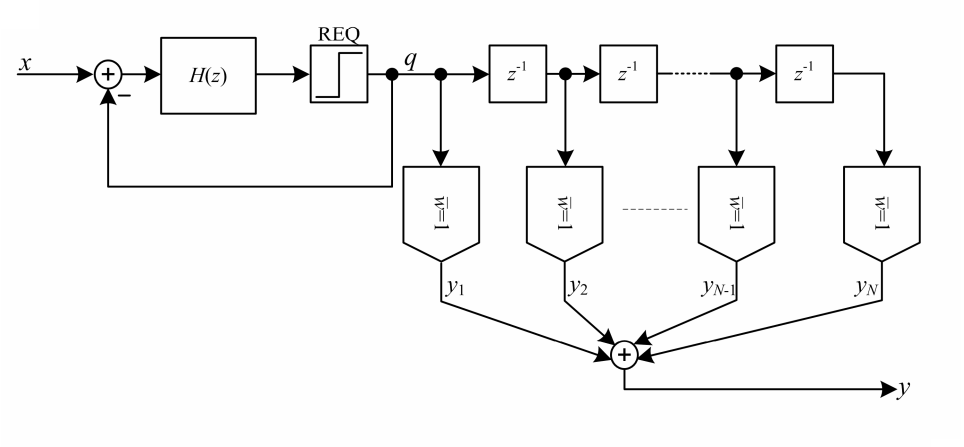


Figure 76: 1-bit DSM REQ with semidigital filtering DAC for multi-level output

With N equally weighted sub-DACs the filtering DAC has a sinc($N\omega$) low-pass response meaning it suppresses out-of-band noise. As long as $L>N$ where L is the OSR, the in-band gain approximates N meaning that y is in practice an N-level signal. Mismatch between the sub-DACs will not lead to distortion as is the case in a regular multi-bit DAC, but will compromise the low-pass function H_{DAC}. Generalized the output is approximately:

$$S_y(\omega) \approx S_x(\omega) \cdot N^2 + \frac{\sigma_{e_q}^2}{2\pi} \cdot \left| NTF(\omega) \cdot H_{DAC}(\omega) \right|^2 . \qquad (156)$$

$H_{DAC}(\omega)$ is the semidigital DAC's frequency response. The SJNR is now approximately:

$$SJNR \approx 10 \cdot \log_{10} \left(\frac{\frac{A^2 \cdot 2^{2B}}{f_{s\_in}^2 \cdot L}}{\left(A^2 \cdot 2^{2B} \cdot \omega_x^2 + \frac{2}{3N^2} \left\| d\left(NTF(\omega) \cdot H_{DAC}(\omega) \right) \right\|_2^2 \right) \sigma_j^2} \right) . \qquad (157)$$

Although it alleviates wideband jitter problems, a filtering DAC does not prevent problems inherent in the 1-bit DSM REQ such as poor stability, limited input-range, idle-tones and noise power modulation. But as should be clear by now; with a high OSR the DSM only needs quite few levels to render REQ quantization noise and related issues negligible. The reason it is desirable to use *many* levels is primarily to alleviate wideband jitter problems. In the fourth paper (Appendix 6), the combination of a few-level DSM REQ and a semidigital filtering DAC to create a *many*-level and relatively jitter-immune output, was explored as an alternative to a many-level DSM REQ with segmented DEM.

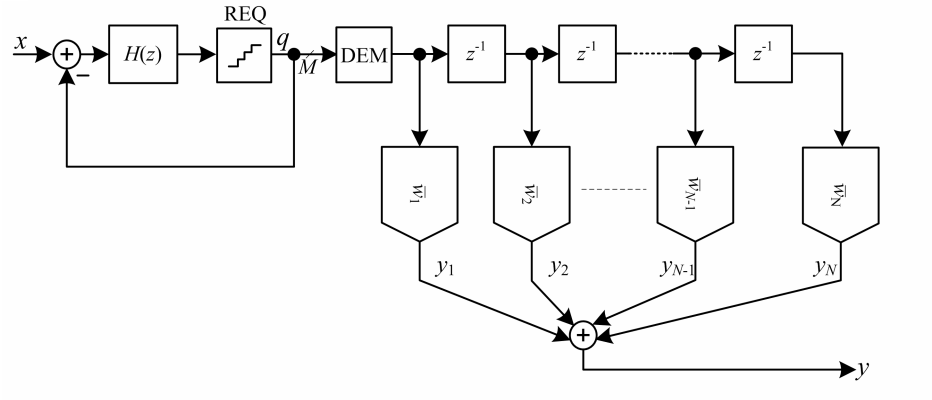


Figure 77: Multi-bit DSM REQ with semidigital filtering DAC

As seen in fig.77 the DSM is now M-level and N sub-DACs are implemented for an effective ($M \cdot N$)-level output signal. Furthermore the DAC weights are generalized since windowed weighting of the sub-DACs gives better out-of-band suppression and thus better SJNR than equal weighting. It is shown in the paper how mismatch compromises the DAC transfer function so that its expected response is:

$$E\left\{\left|H_{DAC}(e^{i\omega})\right|\right\} = \left|H_{ideal}(e^{i\omega})\right| + \sqrt{N} \cdot \sigma_{e_w} . \qquad (158)$$

Where σ_{e_w} is the mismatch error as given in (101). Simulations in the paper show that a DSM REQ with second order DEM and a hann filtering DAC where M=15 and N=17, performs better than a segmented second order DEM DAC with M=255 in the presence of $50ps_{RMS}$ white jitter and 1% RMS mismatch at the 255-level LSB weight. What is the best choice depends on whether mismatch (with DEM) or jitter is expected to be the limiting factor for the final SNR. If jitter noise dominates the semidigital filtering DAC is the better choice, while if mismatch noise dominates the segmented DEM DAC will be the better choice.

5.3.4 Pulse Width Modulating DAC

Pulse width modulation is a way to represent a signal as a two-level waveform. While PCM represents the input as amplitude quantized codes, PWM represents input amplitude samples as corresponding pulse widths in a periodic waveform. PWM was conceptually described in 1933 by Bennett [147], its use in audio amplification suggested in 1965 by Josephson [148]. PWM amplification is attractive because two-level signals facilitate Class-D (switching) amplifiers with very high efficiency [148]. Research has also diverted into digital PCM-PWM conversion for use in DACs [150] and high output power "digital" amplifiers [151].

Figure 78a) shows the conversion of an analog signal to PWM, typically referred to as Natural PWM (NPWM). The PWM waveform is obtained by comparing the input to a

reference carrier in an analog comparator. The carrier is periodic with frequency f_c and one output pulse is generated per period with a width proportional to the input amplitude at the crossing point. Thus NPWM is "time sampling" at the crossing point. To avoid multiple crossing points the slew rate of the carrier must always be higher than the signal. From this it is required that $f_c > f_x \cdot \pi$ for a full-scale sinusoid x. The PWM spectrum will consist of a signal component, a carrier component and modulation products. The input is reconstructed by low-pass filtering after the switching amplifier. For good reconstruction, i.e. high suppression of the carrier component and modulation products, it is common that $f_c >> f_x \cdot \pi$.

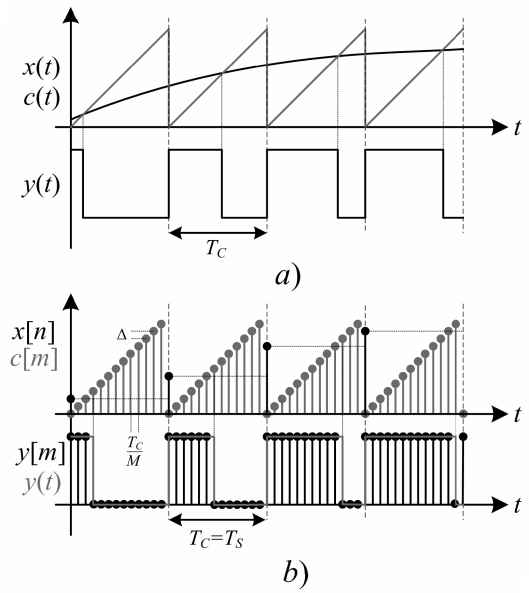


Figure 78: a) Analog PWM modulation b) Digital PCM-PWM conversion

Conversion of PCM to PWM is done similarly, and illustrated in fig.78b). The input sample is held throughout its sample period at one input of a digital comparator. The reference carrier at its other input is generated by a counter. The counter resets at an interval T_s and must count from 0 to 2^B-1 between each reset so that any PCM input sample value can be given a digital PWM representation. This means 1-bit PWM samples are generated at a rate $2^B \cdot f_s$, which for 24-bit 96kHz audio equals 1,600GHz. Analogously, the PWM time resolution corresponding to 24-bit PCM amplitude resolution is 0.6ps. This is obviously not feasible to implement so the input must be requantized first. In "digital amplifiers" for audio it is common to use an 8-bit DSM REQ with an OSR of 8 [152] for a more manageable PWM sample frequency of ~200MHz. The requirement for timing *accuracy* is however unchanged since jitter in the PWM waveform is not shaped by the DSM. Unsurprisingly the jitter susceptibility is comparable to a two-level RZ DAC since PWM is in essence a two-level RZ waveform.

Another major issue in PWM amplifiers with digital modulation is PCM-PWM distortion. Input sample n is held by the comparator from nT_s to $(n+1)T_s$ and resampled at a time instant depending on its value. This happens along a "time grid" given by $T_s/2^B$, but if the resolution is reasonably high it can be approximated as continuous in time. It is then called Uniform PWM. Not unexpectedly, the hold error will fold down into the signal band upon resampling and since the resampling instant is signal dependent it will also cause harmonic distortion [153]. How the hold error changes the PWM pulse width compared to ideal reconstruction is illustrated in fig.79. Since only the value at the crossing point is sampled, it is possible to use algorithms for signal-dependent interpolation to approximate the ideal reconstruction case. Goldberg and Sandler did important early work on this [154] and a comprehensive treatment of several approaches and algorithms is given in Nielsen's PhD-thesis [155].

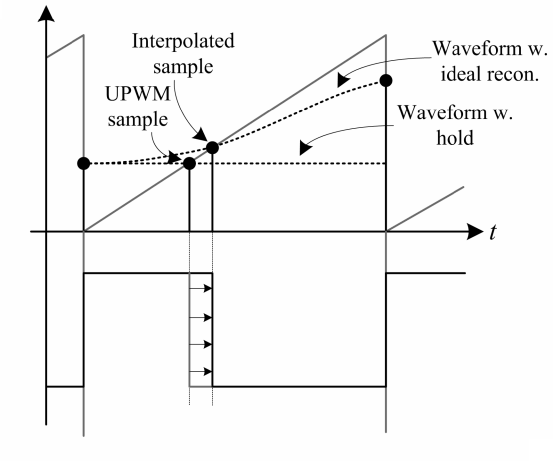


Figure 79: UPWM error

The observant reader will have noted that a switching amplifier can work on any two-level bit-stream, so why not just use a 1-bit DSM REQ which is much more linear than a PCM-PWM conversion? The answer to this is switching losses. With high OSR and high and irregular switching activity, the DSM bit-stream in its basic form is not very suitable to drive high power Class-D amplifiers. Likewise, PWM due to its two-level representation and high jitter susceptibility is not very suitable for high resolution small signal DACs. The research effort into eliminating the weaknesses of both has however led to some convergence, and through new techniques both high-power switching amplifiers based on DSM and hi-res DACs based on PWM have been reported.

"Digital amplifiers" based on 1-bit DSM commonly use quantizer hysteresis [156]-[157] to reduce the switching activity, while some recent hi-res converters have used innovative PWM variations to reduce dynamic errors in multi-bit current DACs. Doorn et al. showed a design using PWM in combination with a semidigital filtering DAC to reduce jitter problems [158]. Each DAC element is fed by a two-level PWM stream making it ISI free, and to avoid PCM-PWM distortion the PWM modulation is done inside the DSM loop. Rueger et al. also showed a solution [159] using several time-interleaved PWM DAC "slices" to control the switching errors and limit the switching activity. The "slices" consisted of semidigital DACs to improve jitter performance.

Reefman et al. showed an ingenious utilization of PWM in a 2003 publication [160], where it is used to eliminate both mismatch noise and ISI while retaining jitter susceptibility at the same level as an ordinary NRZ DAC. In this design each element is PWM modulated and all are used equally regardless of input value. The PWM makes the elements ISI free by ensuring they are switched on and off once every sample, and using them all equally regardless of input value eliminates mismatch distortion. By time-interleaving the PWM modulation within the sample period, the combined output of the current elements equal a normal PCM staircase. This is illustrated in fig.80. Note that the PWM modulation works in modulo fashion so that the active period is rotated when exceeding a sample period.

This algorithm does have a few disadvantages. Firstly, the clock frequency of the PWM logic needs to be $f_s \cdot \mathrm{OSR} \cdot 2^B$ which limits the number of bits in the REQ (and makes it unusable for wide bandwith applications). Also to keep the switching activity constant, the REQ output can only change by ±1 from one sample to the next. Reefman et al. used a limiter inside the DSM loop to force this, but preserving stability then mandates a conservative NTF.

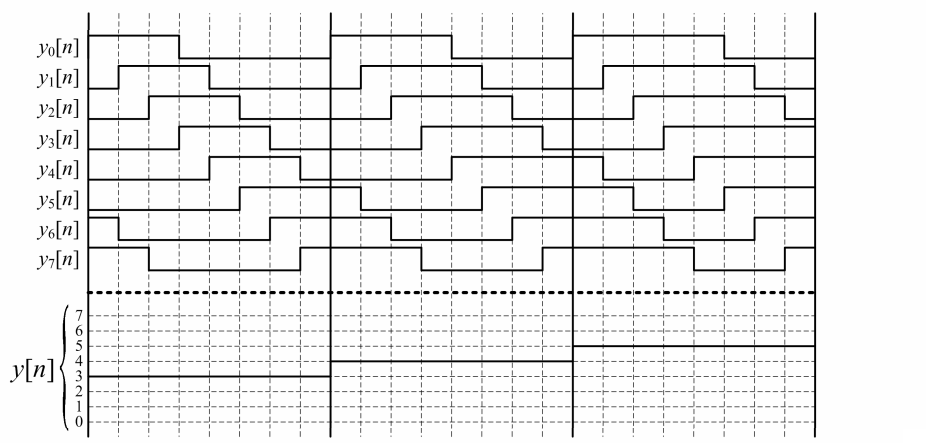


Figure 80: PWM-based algorithm used by Reefman et al. to eliminate mismatch and ISI

A possibility that wasn't explored by Reefman et al. is to use this algorithm in combination with a semidigital filtering DAC. If the DSM REQ and PWM modulated sub-DACs were chosen to be e.g. 5-bit at OSR=128 (for a PWM clock of 400MHz at 96kHz f_{s_in}), and 32 sub-DACs were arranged as a hann-weighted semidigital FIR filter, that would make the DAC highly jitter insensitive and immune to both ISI and element mismatch. For a "super hi-res" implementation this would appear as a very attractive design approach.

Chapter 6

Conclusions and Further Work

6.1 Conclusions

Having digested the five chapters and the overview they give, the reader should be able to assess the challenges and evaluate the results of data conversion design for high resolution audio. It should also have provided the fundament necessary to evaluate the five papers, which deal more specifically with some of the issues that have been presented.

The development of state-of-the-art performance in hi-res audio DACs is illustrated in table 2, listing some key silicon-proven publications. Unfortunately, the relatively low number of published silicon-proven DACs for hi-res audio makes it difficult to produce a survey or performance chart akin to those used for general purpose ADCs [167]. This situation is complicated by published measurements often being made under differing conditions, like signal frequency, amplitude, and frequency weighting. It would in the author's opinion be helpful if designers more strictly adhered to the AES17-1998 measurement standard [168]. The publications in table 2 are selected for having reasonably comparable measurements, and also for illustrating the change of design paradigms: The earliest converter is LPCM and then it moved to (1-bit) DSM, later with switch-cap filtering. In the late 90s multi-bit DSM took over for 1-bit, whereas current-mode DACs superseded switch-cap in the early 2000s. State of the art performance has steadily increased, as has efficiency quantified by the FOM [167]:

$$FOM = \frac{2^{ENOB} \cdot 2f_b}{P} \ . \qquad (159)$$

The ENOB is calculated from the SNDR using the 6dB per bit rule, f_b is the measurement bandwidth and P is the power dissipation in watts. (A) means measurements are A-weighted.

Table 2: Performance development, selected silicon-proven hi-res audio DAC publications

Publication	Topology	SNDR @FS	Power pr.ch.	Meas. bandwidth f_b	ENOB	FOM ($\times 10^9$)
[163] (1986)	16-bit LPCM current-divider	95dB	400mW	20kHz (A)	15.5	4.63
[164] (1987)	1-bit DSM, CT CMOS buffer	90dB	150mW	20kHz	14.7	7.10
[165] (1991)	1-bit DSM, SC DAC	102dB	375mW	20kHz (A)	16.7	11.4
[53] (2000)	Multi-bit DSM, SC DAC	102dB	155mW	20kHz (A)	16.7	27.5
[132] (2000)	Multi-bit DSM, SC DAC	100dB	100mW	20kHz (A)	16.3	32.3
[166] (2000)	Multi-bit DSM, I-DAC	108dB	111mW[9]	20kHz (A)	17.6	71.6
[56] (2001)	Multi-bit DSM, I-DAC	112dB	125mW[10]	20kHz (A)	18.3	103
[160] (2003)	Multi-bit DSM, PWM hyb.I-DAC	>110dB[11]	75mW	20kHz (A)	>18.0	>140

In pursuit for higher resolution still, the designer will have to address all problems dealt with in this thesis. The methods and results presented should make this task easier.

[9] Estimated from data sheet for Texas Instruments PCM1738
[10] Estimated from data sheet for Texas Instruments PCM1792
[11] Limited by resolution of measurement instrument [158]

The work conducted for the first paper was based on extensive simulations and evaluation using Matlab. Four different DSM models were written, including a first order, third order, fifth order and a trellis noise shaping modulator. Their baseband noise power as a function of the input level was simulated by stepping the input and doing a new simulation run for each step. To ensure high enough resolution, each simulation run was 2^{21} samples long and a total of 2^{12} input levels were simulated for each DSM. These included simple fractions of the quantization step to provoke the modulators' idle-tone behaviour. Results show that even if it is high order, a DSM without TPDF dither will have noise-power modulation, but for multi-bit high order modulators it is likely to be negligible compared to circuit noise. Both the third order and the fifth order 1-bit DSMs – the latter being Sony's proposed design for DSD – did however exhibit noise power modulation that will subjectively impede state of the art performance. From these results it is tempting to conclude that SACD or other 1-bit formats will make it very difficult to achieve full transparency, whereas LPCM that in theory is infinitely scalable would be preferable as a raw storage format also in the future.

The research for the second paper was initiated after an e-mail exchange with Peter Kiss, main author of the paper "*Stable High-Order Delta-Sigma DACs*" in TCAS-I [161]. His paper argued for EF modulators being intrinsically more stable than OF modulators, and how a high order multi-bit EF DSM could be designed with guaranteed stability whereas an OF DSM could not. This was found to contradict the conclusions in Kenney and Carley's paper on multi-bit DSM design [123], where the non-overload approach was first introduced. It was found that the cause for the disparaging conclusions between the papers of Kiss and Kenney/Carley was that the former used OF-modulators implemented as mod*N* basic structures (fig.32). Such a DSM does of course not have a unity STF and it was the STF that caused inferior stability. After some further correspondence a paper was written that clarified and extended the non-overload theory, proving the equivalency of OF and EF modulators and now also including truncating quantizers, quantizers with offset and any IIR NTF. The work was again done in Matlab and the model library extended with general model files for any mod*N* DSM, having selectable *N* and quantizer functions.

During an excellent course on delta-sigma at the EPFL in Lausanne Switzerland, Robert Adams in his lecture presentation showcased the advantages of segmented DEM exemplified through his high-end design [140], and argued for a first order SDSM as preferable. A little later a new TCAS paper, "*Multibit Delta-Sigma Modulator with Two-Step Quantization and Segmented DAC*" [162], also discarded the use of a second order SDSM for mandating two extra bits in the compensation DAC. The work done on the non-overload method for the second paper made it clear that it would be applicable here and could be used to design more optimal segmentation modulators. IIR SDSMs were designed and analysed in Matlab and it was found that a very conservative NTF would lessen the complexity penalty by moving to second order to less than half. Using second order segmentation modulators was also found to be hugely advantageous with regards to tones. According to Steensgaard's thesis ([141] pp.174-175), Adams previously argued that tones in the SDSM would not be a problem because its input contained a strong shaped noise component. In his thesis Steensgaard repudiates this claim and simulations done for paper III confirm his reasoning. Matlab models for mismatch DACs with various selectable DEM algorithms and SDSMs were developed in the making of this publication.

The fourth paper was motivated by the difficulty in finding any good documentation for the relationship between the DSM and jitter performance. In numerous publications one can read arguments in favour of multi-bit modulation because of jitter concerns, or that moving from switch-cap to current-steering DACs increases the jitter problem. However it has been difficult to find *quantified* assessments, showing *how* or by *how much* the jitter susceptibility changes when the number of bits, the NTF, or the oversampling ratio is altered. This paper set out as a general study on the relationship between the DSM and jitter errors, but was later

extended to consist of a more general analysis, also evaluating mismatch errors and ISI errors. A range of Matlab models were built for DEM DACs, jittered DACs, and DACs with switching errors, and simplified estimates – that are also shown in chapters 4 and 5 – were developed based on spectral analysis with the additive noise source REQ model. The paper provides estimation methods that should make it easier to predict the distortion caused by circuit non-idealities when designing a DSM converter, or predict e.g. how many bits will be necessary to reach a target SJNR, given a certain amount of jitter. It also clearly shows how advantageous it is to use multi-bit REQ and clarifies common confusions, e.g. surrounding DSM DACs and their susceptibility to different jitter types. The reader should perhaps in particular note how jitter sideband distortion will not be affected by the number of bits or the NTF of the modulator, whereas white jitter noise to a great deal will.

The fifth paper extends on the fourth to investigate some proposals using the simplified estimation methods. The objective was to find a "jitter optimal" DAC within certain complexity constraints. Semidigital filtering DACs have previously been used to improve the jitter performance of 1-bit converters, in this paper it was proposed to combine a multi-bit DSM and DEM with a semidigital multi-bit DAC. The imposed complexity constraint was that the DAC should have 255 levels. An 8-bit DSM REQ followed by a segmented DAC (with a 2.order SDSM) was compared to a 15-level DSM REQ followed by 17 15-level sub-DACs arranged as a semidigital filter. It was confirmed that with proper weighting the semidigital DAC would have significantly better jitter performance than the segmented DAC, with the bonus that the complexity overhead caused by the SDSM and larger DEM network is removed. This topology proved superior with regards to jitter susceptibility, and is recommended to pursue if jitter noise dominates the error budget. In a high-OSR converter for audio it would be a viable approach to achieve very high resolution. In wider bandwidth low-OSR delta-sigma converters it will not be applicable.

6.2 Proposals for Further Work

When this project was initiated, the original intention was to create a chip prototype of a very high resolution audio DAC and base the Ph.D. thesis on measured results. It however became clear after a while that this would be very difficult to achieve. That acknowledgement mainly came from Nordic not having any audio converters in their existing portfolio, and the data converter group not having prior experience with delta-sigma or hi-res converter design. It meant that everything would have had to be made from scratch and with me having no prior design experience it was decided that this plan involved a much too high risk of failure. It was subsequently scrapped and instead the research was directed towards making simulation-based publications, covering topics of interest and where existing published conclusions were lacking or unclear. The idea then was to build a general knowledge base and a simulation model library for future design of audio or delta-sigma converters.

Even though the Nordic data converter team was dissolved, future pursuit or continuation of this project would logically follow that intended path; implementing prototypes and eventually converters based on the knowledge and the methods developed in this thesis and in the papers. If I had one more year to complete my degree, a chip implementation would be the natural step forward and I would propose for a possible successor to embark on this. Then the estimation methods could be confirmed by measurements in addition to simulation results, and the high-level models could be compared with physical implementations of the most promising architectures. It would in particular be interesting to see if the segmented DEM DAC with improved segmentation, or if the combination of a multi-bit DSM and a semidigital FIR DAC, could be used to advance state of the art beyond that shown in Table 2.

Appendix 1

Frequency Analysis

Throughout this thesis, frequency domain simulations are done in the sampled domain. For discrete time signals the DTFT is used to find a continuous spectrum.

$$DTFT\{x\} \overset{def}{\rightarrow} X_s(\omega) = \sum_{n=-\infty}^{\infty} x[n] \cdot e^{-\mathbf{i}\omega n} \quad . \tag{160}$$

The DTFT gives an infinite periodic spectrum. In a real-world simulation scenario, an infinite length sample sequence is generally not available. Assuming the available sample sequence to be of finite length L, its DTFT is:

$$X_s(\omega) = \sum_{n=-0}^{L-1} x[n] \cdot e^{-\mathbf{i}\omega n} \quad . \tag{161}$$

Still the transform is not usable for computer simulation since ω is a continuous variable. The intuitive way to obtain an equivalent fully discrete transform is to sample the DTFT spectrum:

$$X(k) = X_s\left(\omega = \frac{2\pi k}{N}\right) = \sum_{n=0}^{L-1} x[n] \cdot e^{-\mathbf{i}\frac{2\pi kn}{N}} \quad , k = 0,1,\ldots,N-1 \quad . \tag{162}$$

Assuming the available sequence is at least as long as the sample set, i.e. $L \geq N$, the N-point DFT can be defined as:

$$DFT_N\{x\} \overset{def}{\rightarrow} X(k) = \sum_{n=0}^{N-1} x[n] \cdot e^{-\mathbf{i}\frac{2\pi kn}{N}} \quad , k = 0,1,\ldots,N-1 \quad . \tag{163}$$

If $L<N$ the sequence must be zero-padded to do an N-point DFT, this is not the case for any simulations for this thesis. Direct calculation of the DFT is computationally very demanding; its complexity being $O(N^2)$. If N is chosen a power of 2, several algorithms exist to partition the data and speed up the process significantly. These algorithms are generally referred to as Fast Fourier Transforms; a review of FFT algorithms is provided in [169]. An FFT algorithm will typically compute an N-point DFT with $O(N \cdot \log_2 N)$ complexity. Simulations in this thesis and the papers use the Cooley-Tukey FFT algorithm[12] with $N=2^{16}$ unless otherwise noted.

A finite DFT spectrum can have incongruities compared to the real DTFT spectrum of a desired function. If the input signal is a function $x_{in}[n]$ defined in $n \in \langle -\infty, \infty \rangle$, picking a limited sample set of length N to obtain (163) can be rewritten as:

$$x[n] = x_{in}[n] \cdot w[n] \quad , w[n] \overset{def}{=} \begin{cases} 1 \ , \ 0 \leq n \leq N-1 \\ 0 \ , \ \text{otherwise} \end{cases} \quad . \tag{164}$$

[12] Cooley-Tukey is the algorithm used by the default FFT function in Matlab

The input function is multiplied with a rectangular window w of length N, meaning that there is spectral convolution:

$$X(\omega) = X_{in}(\omega) * W(\omega) \quad , W(\omega) = \frac{\sin\left(\frac{\omega N}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} e^{-i\omega\frac{(N-1)}{2}} \quad . \tag{165}$$

The Fourier transform of w is the aliased sinc-function or Dirichlet-kernel. Imagine the input function is a sinusoid $x_{in}[n]=\sin(\omega_x n)$: Then its spectrum is zero everywhere but ω_x. But because the truncated function $x[n]$ is spectrally convolved with the Dirichlet-kernel it will have frequency domain smearing and ringing. When N equidistant samples are taken for the DFT it will have non-zero energy also for other samples than the one closest to ω_x. This is referred to as spectral *leakage*. In fig.81 the result of leakage is illustrated for an example DFT with N=64.

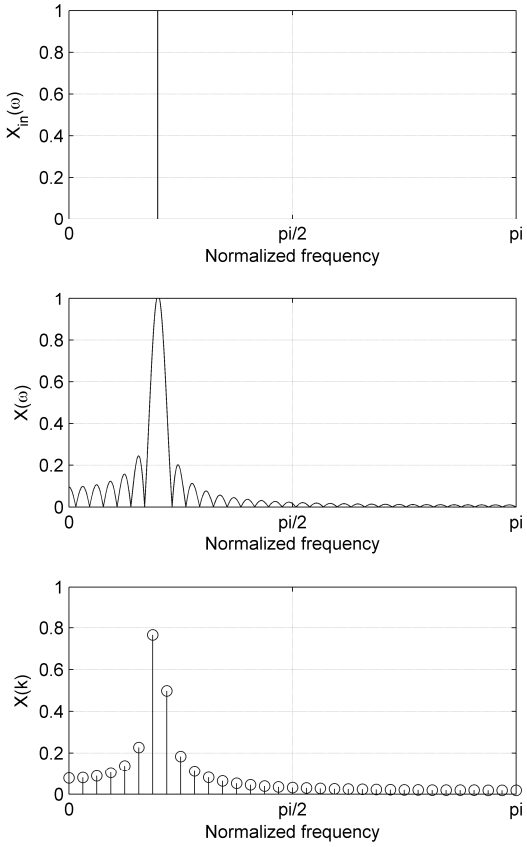


Figure 81: Illustration of DFT spectral leakage

Leakage is less severe with large N, but high resolution SNR simulations are ruined by leakage even if the DFT is extremely long. Because of this *windowing* of the DFT sample set is absolutely necessary. Windowing means to replace the rectangular window defined by picking a sample set of a function with a different window function. A smoother window function will give less ringing by reducing the abrupt end points of the rectangular window. That frequency-domain ringing is complementary to time-domain discontinuities is known from the description of the Gibbs effect [170]. Most simulations in this thesis use the hann-window [171], defined as:

$$w[n] = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) . \quad (166)$$

When the signal is multiplied with the hann-window before doing the DFT, the result for a sinusoid looks like shown in fig.82. As seen the ringing is greatly suppressed and a DFT of reasonable length can now be used for very high resolution simulations. A drawback with windowing is that although side lobes are better attenuated, the main lobe becomes wider. This implies decreased spectral resolution; if there are two distinct tones close in frequency their main lobes from convolution with the window may smear together and it then appears as only one tone in the DFT. Spectral resolution vs. side lobe attenuation is an active trade-off to make when choosing the windowing function. A comparison of the most common windowing functions is found in [172].

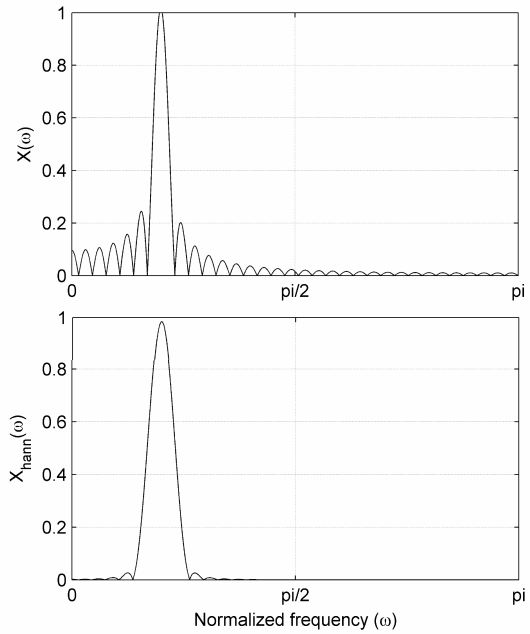


Figure 82: Spectrum of sine multiplied with rectangular (top) and hann (bottom) window

An alternative technique to avoid leakage often used for single-tone simulations is so-called *coherent sampling* [173]. The point with coherent sampling is to set the frequency of an input sinusoid such that the DFT samples correspond exactly to zeros in the convoluted spectrum's side lobes (and the centre of the main lobe). This can be ensured by using an input sinusoid $x_{in}[n]=\sin(\omega_x n)$ with a frequency that fulfils:

$$\omega_x = \frac{2\pi K}{N} \rightarrow f_x = \frac{K}{N} \cdot f_s . \quad (167)$$

K is the integer giving f_x closest to the originally intended input frequency. In this case the time-domain sinusoid has exactly an integer number of cycles and end-point discontinuities are not present. The DFT result is shown in fig.83; it is apparent how $X_{in}(k)$ now will not have any leakage. The result can be confirmed theoretically by correlating the input signal with the DFT basis functions.

It has also been suggested that K should be prime to ensure irreducibility [174]. Then the number of different levels that are excited is maximized, reducing the risk of "hidden" INL errors. This special case of coherent sampling is known as *prime sampling*.

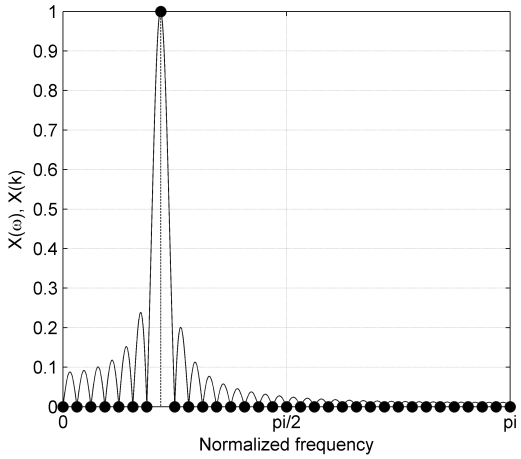


Figure 83: Convoluted spectrum and DFT samples with coherent sampling

A delta-sigma modulator complicates matters somewhat because spectral leakage might impair the results even when coherent sampling is used. The output from a DSM consists of two components, the signal component x and the quantization noise component e_{dsm}. Now even if f_x is chosen coherent and x has no spectral leakage to other DFT bins, the quantization noise e_{dsm} might leak into the signal band. Since the noise is very strongly shaped, especially in high order modulators, leakage from the powerful out-of-band noise may significantly affect the very low in-band noise. This is illustrated in fig.84.

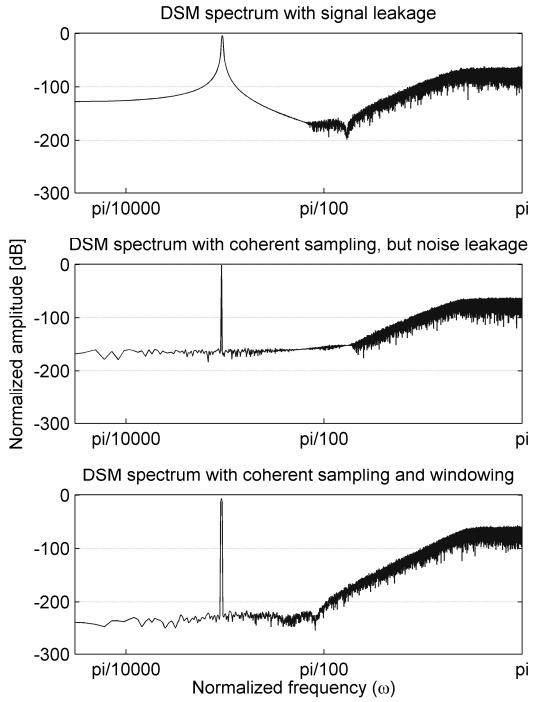


Figure 84: Illustration of signal leakage and noise leakage impairing DSM DFT

In the time-domain noise leakage can be intuitively understood since even though a sinewave has exactly an integer number of cycles within the length of the DFT, *the quantization error superimposed on it* may lead to end point discontinuities. For a high order DSM the quantization error is a shaped noise signal whose time sequence is not possible to derive, and it can't be known how noise leaks in-band. It is therefore *strongly recommended* to use *both coherent sampling and windowing* when doing spectral analysis of a DSM. How this improves DFT resolution by reducing noise leakage is seen in fig.84. Simulations in this thesis use prime sampling and hann windowing. These considerations are treated in more detail in a paper written by the author after initial completion of the Ph.D. studies [175].

Appendix 2

Paper I:

I. Løkken, A. Vinje, T. Sæther, ”Noise Power Modulation in Dithered and Undithered High-Order Sigma-Delta Modulators”, *J. Audio Eng. Soc*., vol.54, no.9, pp.841-854 (2006 Sept.).

© 2006 AES. Reprinted, with permission, from the Journal of the Audio Engineering Society (ISSN 1549-4950)

PAPERS

Noise Power Modulation in Dithered and Undithered High-Order Sigma–Delta Modulators*

IVAR LØKKEN, *AES Student Member,* **ANDERS VINJE, AND TROND SÆTHER,** *AES Associate Member*
(ivar.loekken@iet.ntnu.no) (anders.vinje@iet.ntnu.no) (trond.saether@iet.ntnu.no)

Norwegian University of Science and Technology, Department of Electronics and Telecommunications, 7491 Trondheim, Norway

Dithering and noise power modulation in low- and and high-order oversampled sigma–delta modulators is investigated. Previous publications have presented theoretical analyses of quantizer distortion and noise power modulation in undithered and dithered LPCM quantizers and low-order sigma–delta modulators. However, simulations on practical implementations tend to document only distortion and idle tones and pay no attention to noise power modulation. Consequently there has been some dissension on the requirements for dithering of higher order sigma–delta modulators. Functional simulations of individual error moments in register transfer level models of dithered and undithered sigma–delta modulators are discussed, including first-order and realistic high-order examples. The fundamental difference between 1-bit and multibit quantization is addressed, and two modern techniques for improving 1-bit performance—dynamic dithering and trellis noise shaping—are investigated. Baseband noise power modulation performance is emphasized and shown explicitly for each example since the baseband, although containing only a small portion of the total noise power, is the practical region of interest for oversampled devices. This discussion should provide a pragmatic context in the debate on dither requirements and how to analyze and achieve good noise power modulation performance.

0 INTRODUCTION

The analysis of quantizer error moment dependence was first introduced by Widrow [1]. It is based on the recognition that quantization, rounding an input range $\pm\Delta/2$ to a fixed value, is equivalent to area sampling of the input probability density function (PDF) within that range, also called a quantizer bin. The probability of a given discrete output level y_i is equal to the probability of the input x being within $\pm\Delta/2$ of that level, and the output PDF will be given by

$$f_y(y) = \sum_{i=-\infty}^{\infty} \delta(y - i\Delta) \int_{i\Delta-\Delta/2}^{i\Delta+\Delta/2} f_x(x)\, \mathrm{d}x. \tag{1}$$

This is equivalent to sampling after convoluting the input PDF with a rectangular window of width Δ,

$$f_y(y) = [\Delta \cdot \Pi_\Delta * f_x](y) \sum_i \delta(y - i\Delta). \tag{2}$$

In other words, the input PDF has been area sampled with quantization "frequency" $\phi_q = 1/\Delta$. This is intuitive when looking at Fig. 1, which shows the input and the output in the PDF and signal domains. The Fourier transform of the output PDF, also called characteristic function (CF), is given by

$$\Psi_y(y) = F\{f_y(y)\} - \sum_i \Psi_x(u - i\phi_q)\, \mathrm{sinc}[\Delta(u - i\phi_q)]. \tag{3}$$

Similar to the sampling theorem, Widrow's quantization theorem states that if $\Psi_x(u) = 0$ for $u > \phi_q/2$, the input CF can be fully retrieved from the output CF, and therefore the input PDF from the output PDF. The output CF will merely be a product of the input CF and the CF of the rectangular window (the sinc function). A convolution of the input PDF with a uniform function equals independent white noise added to the input.

Unfortunately no real-world input signals have such a CF as this would require an infinite amplitude span. Those that come closest are large Gaussian-like distributions, as indicated by the findings that led to Bennett's well-known additive noise model [2].

However, one can also use the CF to look at the conditional independence at any given statistical moment of the output. Using the definition of the mth statistical mo-

*Manuscript received 2005 September 29; revised 2006 April 4, May 23, June 8, and June 23.

ment, it can be found by differentiating the characteristic function at the origin,

$$E[Y^m] = \int_{-\infty}^{\infty} y^m f_y(y)\,\mathrm{d}y = \left(\frac{\mathrm{j}}{2\pi}\right)^m \left.\frac{\mathrm{d}^m \Psi_y(u)}{\mathrm{d}u^m}\right|_{u=0}. \quad (4)$$

If a condition is now imposed, namely, that the input CF is of such nature that its mth derivative is zero at all integer multiples of the quantization "frequency" ϕ_q,

$$\left.\frac{\mathrm{d}^m[\Psi_x(u)\,\mathrm{sinc}(\Delta u)]}{\mathrm{d}u^m}\right|_{u=i\phi_q} = 0, \qquad i \neq 0 \quad (5)$$

then it is found, using the previously shown expression for $\Psi_y(u)$ in Eq. (3), that the mth moment of the quantizer output is given by

$$\begin{aligned} E[Y^m] &= \left(\frac{\mathrm{j}}{2\pi}\right)^m \left.\frac{\mathrm{d}^m[\Psi_x(u)\,\mathrm{sinc}(u)]}{\mathrm{d}u^m}\right|_{u=0} \\ &= \left(\frac{\mathrm{j}}{2\pi}\right)^m \sum_{i=0}^{m} \binom{m}{i} \frac{\mathrm{d}^{m-i}\,\mathrm{sinc}(u)}{\mathrm{d}u^{m-i}} \frac{\mathrm{d}^i \Psi_x(u)}{\mathrm{d}u^i} \\ &= \sum_{l=0}^{[m/2]} \binom{m}{2l} \left(\frac{\Delta}{2}\right)^{2l} \frac{E[x^{m-2l}]}{2l+1}. \end{aligned} \quad (6)$$

Solving Eq. (6) one will find that if (and only if) the condition in Eq. (5) holds, the mth derivative of the input CF is zero at all integer multiples of the quantization "frequency." Then the mth moment of $y = x + e$ is equal to the mth moment of x plus a constant, that is, an error that at this moment is input independent and with uniform PDF. Of course, since the source signal x is arbitrary, it is not known whether the requirement in Eq. (5) holds. However, the input of the quantizer can be forced to meet it, regardless of the source, by applying dithering [3]. In this case the quantizer input is given by $w = x + v$, as shown in Fig. 2, with the conditional PDF

$$f_{w|x}(w, x) = f_{(x+v)|x}(w, x). \quad (7)$$

Since the dither is assumed completely independent of the source signal x, it can be found, using some calculus (see [3]), that

$$f_y(y) = [\Delta \cdot \Pi_\Delta * f_v * f_x](y) \sum_k \delta(y - k\Delta) \quad (8)$$

$$\Psi_y(u) = \sum_i \Psi_x(u - i\phi_q)\Psi_v(u - i\phi_q)\,\mathrm{sinc}[\Delta(u - i\phi_a)]. \quad (9)$$

Looking at Eq. (9) it is seen that the same condition as for the undithered case can be applied, but now to the dither signal instead of the input,

$$\left.\frac{\mathrm{d}^m[\Psi_v(u)\,\mathrm{sinc}(u)]}{\mathrm{d}u^m}\right|_{u=i\phi_q} = 0, \qquad i \neq 0. \quad (10)$$

If this holds, the mth output moment will be given by

$$\begin{aligned} E[Y_m] &= \left(\frac{\mathrm{j}}{2\pi}\right)^m \left.\frac{\mathrm{d}^m[\Psi_v(u)\,\Psi_x(u)\,\mathrm{sinc}(u)]}{\mathrm{d}u^m}\right|_{u=0} \\ &= \left(\frac{\mathrm{j}}{2\pi}\right)^m \sum_{i=0}^{m} \binom{m}{i} \left.\frac{\mathrm{d}^i[\Psi_v(u)\,\mathrm{sinc}(u)]}{\mathrm{d}u^i} \frac{\mathrm{d}^{m-i}\Psi_x(u)}{\mathrm{d}u^{m-i}}\right|_{u=0} \\ &= \sum_{r=0}^{m} \binom{m}{r} \sum_{l=0}^{[r/2]} \binom{r}{2l} \left(\frac{\Delta}{2}\right)^{2l} \frac{E[v^{r-2l}]}{2l+1} E[x^{m-r}]. \end{aligned} \quad (11)$$

Since the dither and input signals are assumed statistically independent, $E[v^k]\,E[x^l] = 0$ for any k and l. Solving Eq. (11) it is seen that if the condition in Eq. (10) holds, the mth moment of y equals the sum of the mth moment of x, the mth moment of v, and a constant.

In other words, the error for this moment will be additive and uniform. A dither signal consisting of the sum of N independent rectangular PDF (RPDF) sources of width $\pm\Delta/2$, often called Nth-order dither source, will have the CF

$$\Psi_v(u) = \mathrm{sinc}(\Delta u)^N. \quad (12)$$

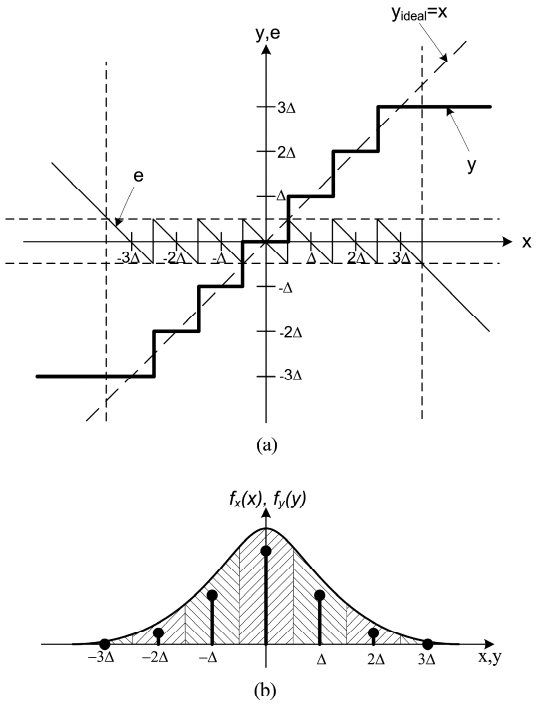


Fig. 1. Midrerad quantizer. (a) In amplitude domain. (b) In PDF domain.

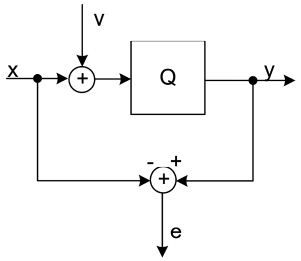


Fig. 2. Dithered quantizer.

It will thus meet the condition of Eq. (10) for the N first derivatives and make the N first error moments uniform and input independent.

In audio we are not interested in harmonic distortion or noise power modulation. Consequently the error expectance value and variance, or the first and second error moments, should be uniform. Higher order moments have been shown to be inaudible [4]. Inserted in Eq. (11), Nth-order dither for $N \geq 2$ gives

$$E[Y] = E[X] \tag{13}$$

$$E[Y^2] = E[X^2] + \frac{(N+1)\Delta^2}{12}. \tag{14}$$

Second-order dither, also called triangular PDF (TPDF) dither because of its distribution shape, is regarded as optimal for quantizers used in audio applications since it renders the two first error moments uniform and produces less additional noise power than higher order dither.

1 QUANTIZATION AND DITHERING IN SIGMA–DELTA MODULATORS

1.1 General Analysis

The effect of quantization and dithering in sigma–delta modulators (SDMs) has been subject to much debate. Early papers claimed SDMs to be self-dithering [5]–[7] while others have rejected this claim vigorously [8]–[14].

The SDM, with its signal flow shown in Fig. 3, is characterized by the differential equation

$$Y(z) = \frac{H(z)}{H(z)+1} X(z) - \frac{1}{H(z)+1}[V(z) + Q_w(z)] \tag{15}$$

where Q_w denotes the quantizer error, a deterministic function of the quantizer input w (see Fig. 3).

In a normal SDM the loop filter $H(z)$ is designed with very high gain at dc, that is, $H(1) \to \infty$. Consequently the second term in Eq. (15) goes toward zero for dc input. This means that average output will equal average input, and thus the first error moment will always be zero.

This observation led to the early assumption of SDMs as being self-dithering. Furthermore the quantizer input in SDMs is largely random. This then also applies to the quantizer error being fed back in the loop. Since the quantizer error is distributed within $\pm\Delta/2$, the feedback was assumed to act like dithering of the first-order kind. With additional dither v, both q and v are present in the feedback signal, and a new v is added to the quantizer input as well. Hence if v is a first-order RPDF dither signal, the total dithering would then be of the third-order kind. It was thus concluded that an SDN with Nth-order dither would have error moments equal to an LPCM quantizer with $(2N + 1)$th-order dither [6].

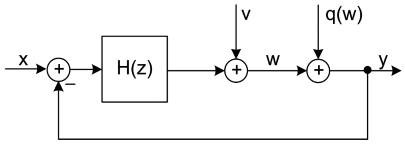


Fig. 3. General sigma–delta modulator.

This argument was later strongly repudiated [14], since it was pointed out that even if q looks random, it is not statistically independent of either v or x. Thus it is wrong to sum q and v into a higher order dither source. This was substantiated using the simple example $x = 0$. Even with RPDF dither the quantizer can never be triggered in this case; y is then necessarily always 0, leading to $q = 0$ and consequently $e = 0$. This shows that the second moment of e is not rendered conditionally independent of the input, even when using RPDF dither. A more rigorous mathematical proof of internal signal dependence in SDMs is found in the extensive mathematical work by Wannamaker [15], [16], where it is shown that to have guaranteed conditional independence in the mth moment of the quantization error, the dither input v in an SDM, will also have to be of the mth-order kind.

1.2 The 1-bit Case

The analysis of 1-bit quantizers differs somewhat from that of multibit ones, and for general SDMs it can be simplified significantly for dc statistics [6].

As has already been shown, the average output in a normal SDM equals the average input, regardless of whether or not the quantizer is dithered. Since the output has only two levels, −1 and 1, its PDF will be two Dirac pulses at exactly those values. Furthermore, since the average output equals the average input, the output for any dc input value x must obey the following condition:

$$P(y = 1) - P(y = -1) = x. \tag{16}$$

We also know that the sum of all possible probabilities is always equal to 1. Hence,

$$P(y = 1) + P(y = -1) = 1. \tag{17}$$

Combining these two requirements, the output PDF is found,

$$f_y(y) = \delta(y-1)\frac{x+1}{2} + \delta(y+1)\frac{1-x}{2}. \tag{18}$$

Since $e = y - x$, the error PDF has to be

$$f_e(e) = \delta(e+x-1)\frac{x+1}{2} + \delta(e+x+1)\frac{1-x}{2}. \tag{19}$$

Then the error moments can be calculated from the expression for the PDF,

$$E[e] = 0 \tag{20}$$

$$E[e^2] = 1 - x^2 \tag{21}$$

$$E[e^3] = 2x(1 - x^2). \tag{22}$$

For the general 1-bit SDM, this output PDF and these error moments are asymptotically correct regardless of whether or not the quantizer is dithered, meaning that every modulator of this kind will have the same noise power modulation. The argument for dithering is consequently fundamentally different in a 1-bit modulator than in a multibit one. In the 1-bit case the dither cannot change the input dependence of individual error moments, but will

only affect the dynamic behavior of the modulator. It can reduce the presence of idle tones, but noise power modulation is unavoidable. In a multibit modulator dithering can prevent tones as well as noise power modulation.

2 SIMULATIONS AND RESULTS

In this chapter the simulations and the results will be presented. First simulations are shown for an LPCM quantizer, which will serve to introduce the analysis and the results to the reader. Next the SDM and the baseband SDM analysis are introduced through a first-order example, and simulation results are discussed. Finally, results are provided and analyzed for three different examples of realistic high-order SDMs.

2.1 LPCM Quantization

This section shows functional simulations done on a MATLAB model of a midrerad quantizer. For a range of different input levels the mth error moment is estimated by taking a long-term average of the measured e^m. The error moment is then calculated as a function of the input level to see whether it is uniform or input modulated.

The simulations are done with no dithering, RPDF dithering, and TPDF dithering, and the results are shown in Fig. 4. As expected, it can be seen that the first error moment, or the mean, is highly input dependent. This leads to harmonic distortion on the output. RPDF dither renders the first error moment input independent, as expected from the theoretical derivation, and thus eliminates harmonic distortion. Only TPDF dither also eliminates noise power modulation. The penalty is higher average noise power, as indicated by Eq. (14).

The simulation results are in conformance with theoretical expressions shown previously for these error moments [3]. The slight "noise" seen on the otherwise flat plots is due to a limited length in the simulation runs, which gives some moment estimation uncertainty. To be perfectly RPDF or TPDF distributed, the dither sequence would have to be of infinite length, which is not possible in a functional simulation. The run length is chosen so that the estimation noise does not mask the results, and in the LPCM case, it is 2^{16} samples per dc input level. The number of dc input levels simulated is 2^{11}, and they are distributed so that the input matches all integer multiples of $B\Delta/2048$ within each and every quantizer bin, B being the number of bins in the quantizer.

2.2 First-Order SDM

This section will review simulated error statistics for a traditional first-order SDM, implemented as shown in Fig. 5. Since the first error moment will converge to zero and error moments of 3 or above are considered irrelevant for audio, only the second error moment, or noise power modulation, is calculated. Simulations are done both for the multibit and the 1-bit case. In addition, noise power modulation is shown for the baseband.

2.2.1 Noise Power Modulation, Multibit

The results from error moment simulations on a 4-bit first-order SDM are shown in Fig. 6. As is seen, the nondithered SDM indeed has noise power modulation patterns

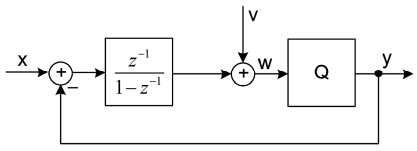


Fig. 5. First-order SDM used for simulations.

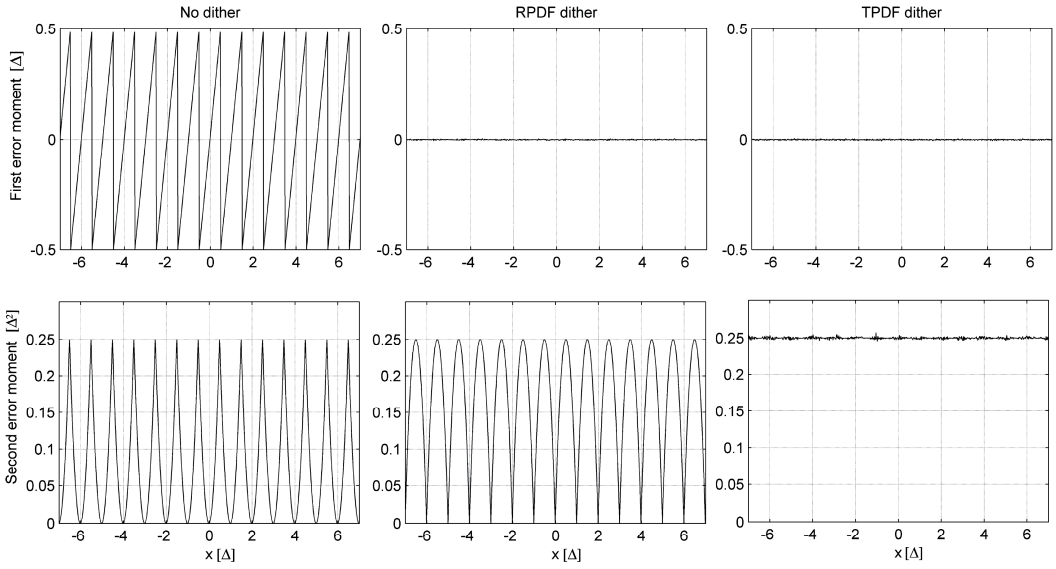


Fig. 4. Simulated error moments, LPCM quantizer.

similar to the RPDF-dithered LPCM quantizer. However, an important difference is that for a dc input the output error from a first-order SDM is periodic instead of white. It is well known that this tonal behavior will make the modulator unusable without dithering.

The second error moment in the first-order SDM with *N*th-order dither is similar to the LPCM quantizer with (2*N* + 1)th-order dither, but as the simulations show, there are deviations. As expected, the noise power is zero at the center of each bin. At these points the quantizer output equals the input, and the loop difference is zero. The RPDF dither amplitude of Δ/2 is not sufficient to trigger quantizer transitions at these exact points. As Fig. 6 shows, with RPDF dither there are also dips in the noise power at other fractions of the bin width. Both at ±Δ/2 and at multiples of Δ/4 visible artifacts can be seen. Fig. 7 shows a closer view of the noise modulation within a bin. The noise power at multiples of Δ/2 goes down from $\Delta^2/3$ to $\Delta^2/4$, and at the quarter of the bin width it dips to around $5\Delta^2/16$.

We can also see in Fig. 6 that the noise power decreases toward the end of the input range. This happens because the quantizer overloads, producing an overweight of maximum-level output values. It leads to a skewed error pattern where the power will be dependent on the distance from the maximum level to *x*.

Overall the first-order SDM with *N*th order dither behaves quite similar to a LPCM quantizer using (2*N* + 1)th-order dither, except where the position of the input signal within a quantizer bin is a simple fraction of the bin width. For no noise power modulation the first-order SDM has to be TPDF dithered.

An SDM is usually a highly oversampled device, and it is the baseband that is mainly of interest in practical applications. To enable estimations of noise power modulation in the baseband, the power spectral density is calculated for each measurement before being integrated over the region of interest to find its total power. 32 times oversampling is assumed, and consequently the baseband is defined as $\pm f_s/64$. The baseband noise power is shown in decibels compared to the power of a full (stable)-scale sine-wave input (dBfs).

We can see from the baseband noise power simulations in Fig. 8 that the undithered noise power is found to be very peaky. As mentioned previously, the first-order modulator is highly tonal, and the noise power within the baseband depends largely on whether or not the tones generated will fall within it for the given dc input. As can be seen from Figs. 8 and 9, the undithered modulator noise power consequently varies greatly with the input, peaking at levels that produce low-frequency cycles.

These levels are, as expected, close to zero and adjacent to simple fractions of Δ. Fig. 8 shows the noise power over the entire input range, whereas Fig. 9 shows a zoomed-in version of the undithered second moment to illustrate

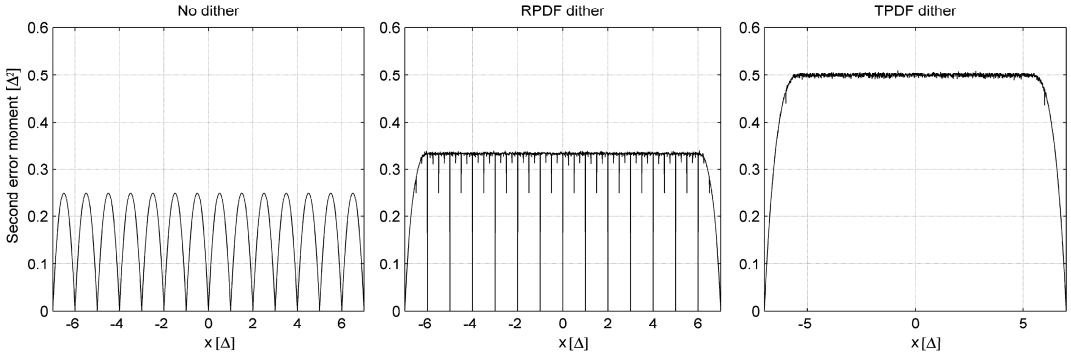


Fig. 6. Simulated noise power modulation, first-order 4-bit SDM.

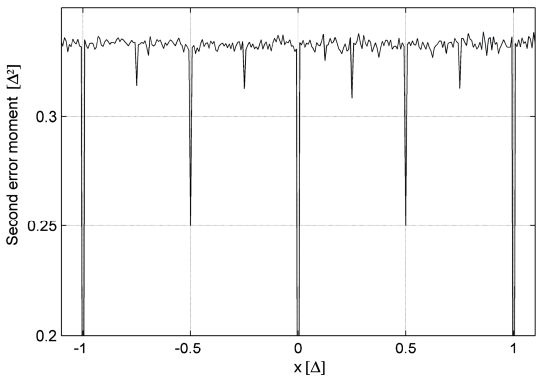


Fig. 7. RPDF-dithered first-order 4-bit SDM; noise power modulation within quantizer bin.

more clearly the position of the noise power peaks. The largest peaks are found around zero as low-level dc input will produce low-frequency tones.

With dithering the tonal behavior is eliminated, and it is seen that although the amplitude has changed, the baseband error moments have the same basic structure as the total error. The exception is when quantizer saturation occurs, in which case it will output limit values more and more predominantly as the overload increases, shifting the error down in frequency and into the baseband.

We have shown that in the dithered case the basic structure of the input dependence of error moments is the same in the baseband as for the entire modulator. Again, for no noise power modulation it is necessary with TPDF dithering. It should also be noted that quantizer saturation is a source of increasing baseband noise power, and that the modulator should be designed so that this is minimized.

2.2.2 Noise Power Modulation, 1 bit

The functional simulations of the 1-bit first-order modulator, shown in Fig. 10, confirms the analysis in Section 1.2. The total noise power is the same regardless of dithering and follows the theoretical expression in Eq. (21) in all cases.

Since full dither will overload a 1-bit quantizer constantly and is not realistic, low-level and high-level RPDF dither of width 0.2Δ and 0.5Δ, respectively, is applied in this case. Although the high-level dither will overload the quantizer for most inputs, a first-order modulator is inherently stable and will track the input nonetheless. This is also confirmed by simulation since the noise power following the theoretical expression requires the average output to equal the average input. Simulation of the 1-bit first-order modulator confirms the theoretical analysis in Section 1.2, that is, the total noise power modulation is not affected by dithering of the quantizer.

In the baseband the 1-bit modulator exhibits behavior quite similar to the otherwise identical multibit version. The dc levels creating tones within the baseband are seen from the bottom plots of Fig. 10 to produce large peaks in the baseband noise power. As expected these levels are found close to zero and adjacent to simple fractions of Δ. The largest peaks are found around zero as a very low-level dc input will produce very low-frequency tones. There are also large peaks around $\pm\Delta/3$, $\pm\Delta/2$, and $\pm 2\Delta/3$.

The low-level dither does reduce the number of occurrences of baseband tones, but for some input levels they are still strong, peaking at almost 20 dB higher than the nominal noise power. The modulator can be expected to have significant tonal behavior, even with this dither. With high-level dithering most limit cycles are removed, with peaks now only occurring for inputs close to zero, $\pm\Delta/3$, and $\pm\Delta/2$. Unlike the total noise power, the nominal baseband noise power is largely flat, but increases for high levels due to increasing quantizer overload. More dither will strengthen this behavior further.

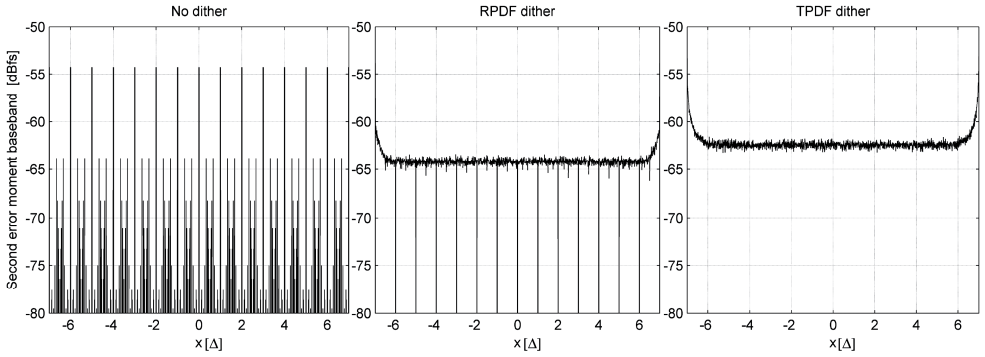


Fig. 8. Simulated baseband second error moment, first-order 4-bit SDM.

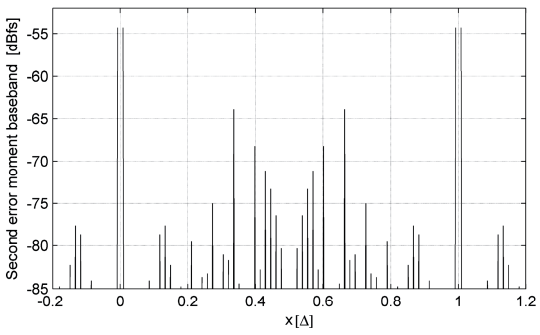


Fig. 9. Undithered first-order 4-bit SDM; baseband error power within quantizer bin.

These simulations have shown that the 1-bit modulator will always exhibit noise power modulation, even in the baseband. Using no dither or low-level dither causes insufficient tonal performance, while high-level dither will increase the overall noise power and the baseband noise power modulation due to quantizer overload effects. The dithering should consequently be chosen based on the requirements for sufficient tonal performance. Increasing the dither level beyond that will lead to more quantizer overload and more baseband noise power modulation.

2.3 Higher Order SDMs

In this section the investigation is extended to simulations of real higher order implemented modulators. They include a multibit third-order SDM used in a commercial audio digital-to-analog converter [17], a fifth-order 1-bit SDM suggested for the Super Audio CD (SACD) format [18], and multibit and trellis versions of the same fifth-order SDM. In the following the focus will again be on noise power modulation only, as the first error moment always converges toward zero while the third and higher moments are of little interest in audio.

In the high-order SDM simulations the simulator run length is increased to 2^{21} samples per period to prevent finite-length dither sequences from adding distribution noise to the moment estimations, obscuring the modulation patterns in the very low-amplitude baseband. It is also necessary to ensure sufficient resolution for the power spectral density calculations.

2.3.1 Fifth-Order 1-bit Modulator for SACD Applications

This high-order modulator is based on a fifth-order feedforward design suggested for 1-bit, 64 times oversampling SACD applications [18]. The block diagram is shown in Fig. 11. With a 1-bit quantizer it is stable for

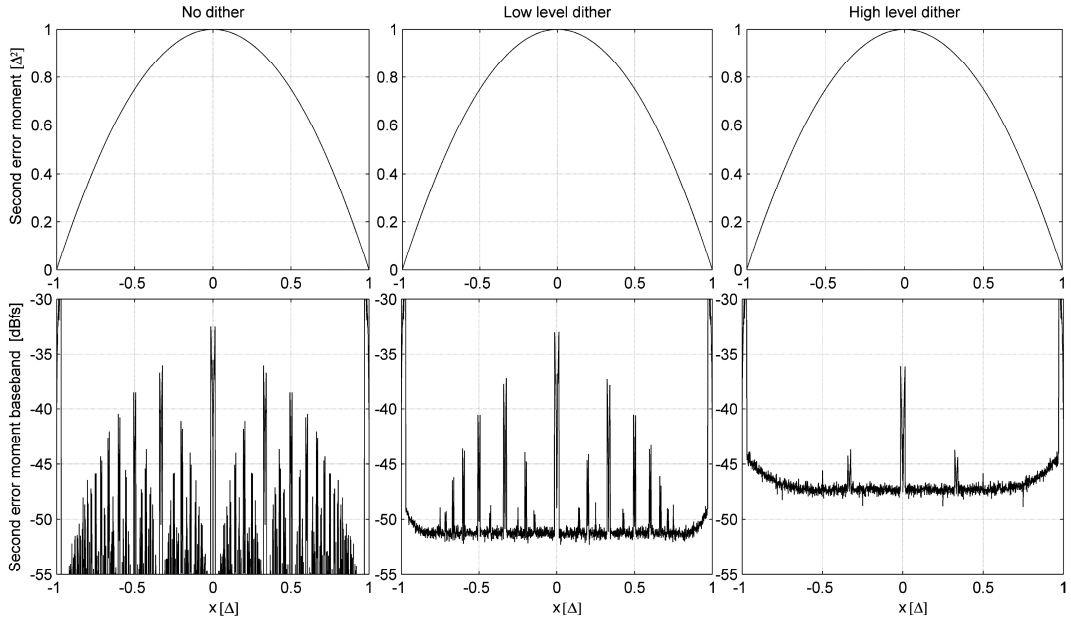


Fig. 10. Simulated noise power modulation, first-order 1-bit SDM.

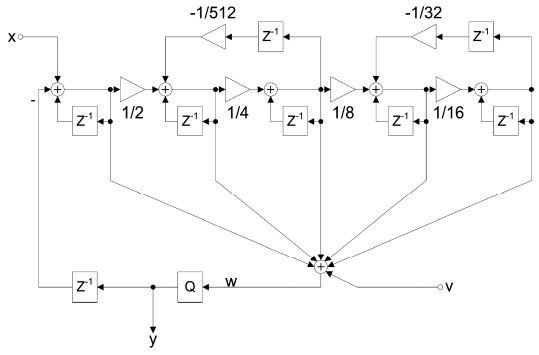


Fig. 11. Fifth-order feedforward SDM.

input up to ±Δ/2 and has 120-dB baseband SNR. RPDF dithering of amplitude 0.25Δ has been shown to break up any visible limit cycles and is consequently recommended by the author [18]. Simulations are performed without dithering as well as with the recommended dither.

As derived earlier, the 1-bit total error moments will be the same regardless of dither and modulator topology, as long as the modulator follows the basic requirement that the average output equal the average input. The simulation results confirm this. In the stable input range the second error moment, as seen in Fig. 12, follows the exact $(1 - x^2)$ pattern of the first-order version, again confirming the analysis in Section 1.2 as being generally valid.

Above this range, however, the modulator becomes unstable, and the output will no longer track the input. It is seen that the usable input range is a bit more than ±Δ/2, as shown also in Takahashi and Nishio [18]. The ±Δ/2 level corresponds to 0-dB_{SACD} or maximum input with a direct stream digital (DSD)-conform source. Applying 0.25Δ RPDF dither to the input, the range of stable operation is limited to a bit more than ±0.3Δ or −4 dB_{SACD}. The simulations confirm the analysis of 1-bit modulators, and in the stable range the error moments correspond to the theoretical findings. The modulator will have the same noise power modulation, regardless of whether or not dithering is applied at the input of the quantizer.

The lower graphs in Fig. 12 show the baseband part of the simulated second error moment. Since the oversampling ratio is 64, this is found as the integrated noise power density in the region of $\pm f_s/128$. The noise power modulation shows an increase for larger input levels due to quantizer overload. A high-order 1-bit modulator will always have some degree of quantizer overload regardless of the dc input. Consequently the gradual shifting of error power toward lower frequencies will increase continuously from zero input and upward. Fig. 13, which displays the baseband noise power with dither as well as a histogram of the relative occurrence rate of quantizer overload, shows how explicit this relationship is.

Some peaks can be seen in the undithered case, indicating that there are some dc levels that will cause the modulator to produce low-frequency tones, although at very low levels. The total baseband error power increases to around −116-dBfs for the most tonal cases, corresponding well to the −120-dBfs distortion components reported by Takahashi and Nishio [18]. The dithering can be seen from the lower part of Fig. 12 to remove it, reduce the tones to virtually immeasurable levels. This is the same as found by Takahashi and Nishio [18] whose spectral plots reveal 0.25Δ RPDF dither to completely eliminate any visible idle tones. In this case the baseband noise power modulation is dominated by a rise in noise power for high input levels because of the increasing quantizer overload. The noise power varies by around 2 dB over the stable range. Also, it should be noted that the nominal dynamic range is around 5 dB less than undithered, because of the reduced stable input range as well as the higher noise power due to the dithering.

This tradeoff between tone performance and quantizer overload, which also compromises modulator stability, has been addressed by Norsworthy [19] and Angus [20], who

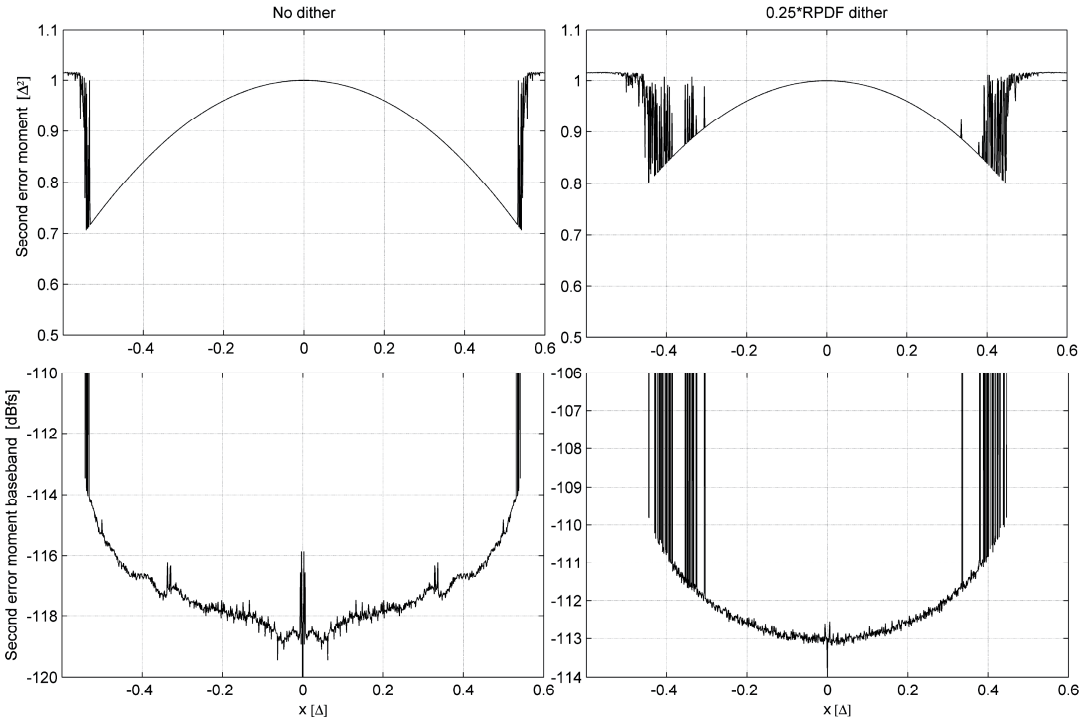


Fig. 12. Simulated noise power modulation, fifth-order 1-bit SDM.

suggest various forms of dynamic dithering. Fig. 14 shows the results for the 1-bit fifth-order modulator in Fig. 11, dithered with Norsworthy dynamic dithering [19]. This algorithm scales the dither depending on the input signal amplitude. For low input levels, which usually excite the strongest baseband tones, the dither level is high. For high input levels the dither is reduced in order to reduce overload and improve stability. The general algorithm [19] is given by

$$v[n] = d[n](1 - |x|^{\alpha}) \tag{23}$$

where $v[n]$ is the dither applied to the input and $d[n]$ is the unprocessed dither source. For this 1-bit modulator α is set to 0.5. As can be seen from Fig. 14, the stable input range is increased to almost 0.5, or 0 dB_{SACD}, almost as good as without dither. At the same time the tonal performance is comparable to the statistically dithered version. The noise power modulation and overall dynamic range are also improved significantly as compared to the statistically dithered case because of an increasingly stable input range, and since the variation in dither power actually counters the variation in error power due to quantizer overload. Some concern has been stated over this algorithm since the dither power is signal dependent [20]. But in the 1-bit case the total error power is given by Eq. (21) as long as it is stable, and the baseband noise power is still dominated by quantizer overload. The simulations clearly show its noise power modulation performance as well as the dynamic range being improved.

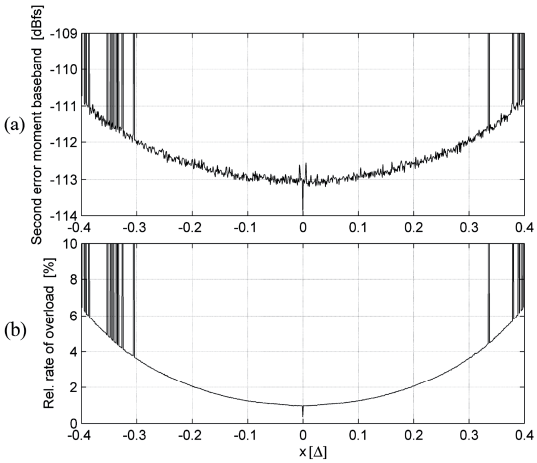


Fig. 13. (a) Baseband noise power modulation. (b) Relative rate of quantizer overload.

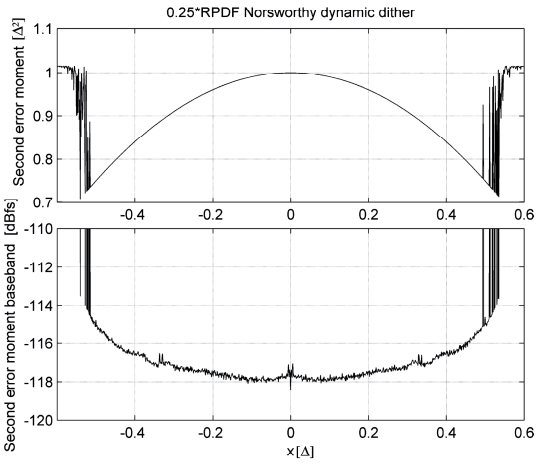


Fig. 14. Simulated noise power modulation, fifth-order 1-bit SDM with Norsworthy dynamic dithering.

In conclusion, the simulations found much the same results for the 1-bit fifth-order modulator as for the 1-bit first-order case. When stable, the total noise power modulation will be as derived in Section 1.2, regardless of whether or not the SDM is dithered. In the baseband the noise power modulation will be given by the degree of in-band tones, as well as by shifting the noise power to lower frequencies due to quantizer overload. Dithering will serve to eliminate the tones, but it will also increase the baseband noise power because of dither noise and more overloading. Dynamic dithering has proved to be an efficient method to minimize these problems.

2.3.2 Third-Order 3-bit Modulator for High-Performance Audio DAC

In this section simulations done on a higher order multibit modulator used in a commercially available audio DAC are described [17]. The modulator is a feedback structure with a seven-level (3-bit) midread quantizer. The suggested dithering is a $(1 - z^{-1})$-filtered RPDF dither of width 0.156Δ, which will be used in the simulations together with an undithered version of the same modulator. The modulator is designed for 64 times oversampling, is reported to have 111-dB dynamic range undithered and 120-dB dithered, and has a signal flow as seen in Fig. 15.

The simulations are done with the same premises as for the fifth-order 1-bit modulator. Fig. 16 shows the output noise power versus input level for undithered and dithered modulators. In addition it shows their baseband components. The modulator is stable for ±2.7Δ undithered and ±2.6Δ when dithered. As can be seen clearly, the modulator is not self-dithering in terms of noise power modulation. In addition the baseband simulations show that in the undithered case it has low-frequency tones for many different input values, although all with a total power of less than −108 dBfs. Even with the recommended dithering there are tones present when the input is adjacent to an integer multiple of Δ, but the dithering does reduce the

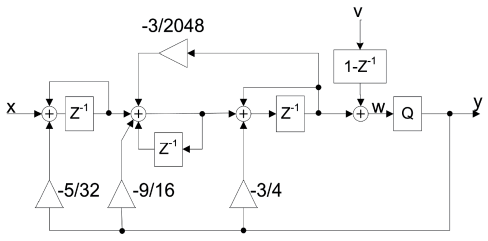


Fig. 15. Third-order 3-bit SDM.

tonal behavior significantly. However, the noise power modulation is not reduced much. Even without including the tones, it is seen that the noise level in the baseband will vary by more than 5 dB, from −117 to −112 dBfs, despite the dithering. The total noise power varies from 0.22 to $0.31\Delta^2$.

This shows the third-order modulator to be insufficiently dithered with regard to noise power modulation. However, as it is designed, full TPDF dithering of maximum amplitude ±Δ would not allow a usable stable input range. If noise power modulation is to be avoided, the modulator design must be with a larger input range for which stability is maintained, allowing full TPDF dithering to be applied to the quantizer. These simulations have shown that a high-order multibit modulator is not self-dithering and that low-level dither, although eliminating much of the tone problem, will not necessarily prevent significant noise power modulation of the entire noise spectrum or in the baseband.

2.3.3 Fifth-Order 4-bit Modulator

The modulator in Section 2.3.1 is designed to be stable for a reasonable input range with only 1-bit quantization. With a multibit quantizer it is not an optimal modulator as the noise transfer function (NTF) is overly conservative for such a design. Still, it is an interesting case for simulations. First, it can be seen whether the more complex loop behavior of such a high-order modulator makes it more self-dithering than the third-order one. Furthermore the conservative NTF allows full-scale RPDF or even TPDF dither to be added while still retaining a large stable input range. Consequently the same SDM shown in Fig. 11 is analyzed, but now with a 4-bit (15-level) quantizer and full ±Δ/2 RPDF as well as ±Δ TPDF dithering.

As is seen from the noise power simulation results in Fig. 17, the undithered modulator definitely has a distinctively input-dependent noise power pattern. The total noise power varies from 0.38 to $0.46\Delta^2$. In the baseband it varies by around 1 dB and has 2.5-dB peaks due to low-frequency limit cycles at input levels around zero and multiples of Δ. Even with full RPDF dithering the noise modulation is clearly visible, with the total noise power varying from 0.58 to $0.65\Delta^2$, but the baseband noise power varies by less than 0.5 dB around a very low nominal level of almost −142 dBfs. When using full TPDF dither the modulator exhibits no noise power modulation.

The results again refute the assumption that a multibit SDM with *N*th-order dither is equivalent to an LPCM quantizer with $(2N + 1)$th-order dither. They also confirm that for constant noise power, even very high-order multibit SDMs need full TPDF dither. However, with such a high-order DSM both the absolute noise power level and its variations are so low that it is more of theoretical than practical interest.

2.3.4 Trellis Noise-Shaping 1-bit Modulator

The trellis noise-shaping (TNS) modulator is a variant of look-ahead sigma–delta modulation, first introduced by Kato [21]. The look-ahead SDM principle is shown in Fig. 18. The input signal x is compared to a range of vectors with 1-bit representation $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N]$. For each vector the error is weighted by a filter $H(z)$, and a cost function c is evaluated. This cost function can be chosen freely, but usually the mean square error, or error power, is preferred.

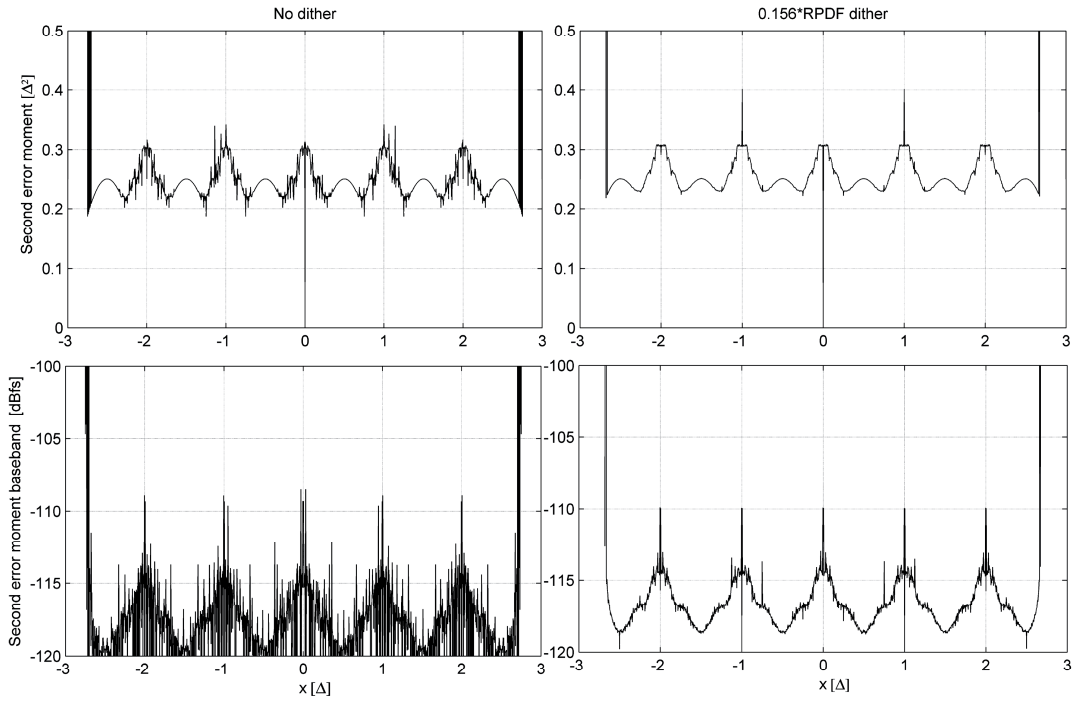


Fig. 16. Simulated noise power modulation, third-order 3-bit SDM.

The vector $\boldsymbol{x}_n$ that produces the smallest cost is chosen to be the output. Since the error is directly weighted by $H(z)$, the corresponding NTF is given by

$$\mathrm{NTF}(z) = \frac{1}{H(z)}. \tag{24}$$

Ideally, to find the global minimum the vector $\boldsymbol{x}_n$ should be the single combination of ones and zeros that produces the least cost for the entire sequence of input data. This is of course not feasible to implement, since one would then have to try every possible combination of ones and zeros corresponding to the length of the entire data set. Rather, the vector is developed in real time using the Viterbi algorithm [22]. The Viterbi algorithm searches through a trellis of limited-length candidate sequences, with the length often referred to as the trellis order, finding the path that minimizes the cumulative cost as it moves along. It then backtracks the path to produce the resulting output vector. The path converges in time, meaning that all candidate sequences are likely to have originated from a shared sequence when backtracking a large number of samples. The backtracking depth can thus be limited to avoid excessive latency from input to output. A disadvantage of the TNS modulator is its complexity, which increases exponentially with the trellis order. This has led to recent research efforts into more efficient implementations [23]–[25]. For a more detailed mathematical description of the TNS modulator the reader is referred to [21]–[25].

To provide easily comparable results, the noise power modulation simulations have been performed on a TNS modulator using the same NTF as the fifth-order SDM analyzed previously in Sections 2.3.1 and 2.3.3. For the simulations the trellis order has been set to 4, and the backtracking depth to 2^{12} samples. The implementation is based on the full trellis algorithm as originally described by Kato [21], and the cost function is the mean square of the loop-filter output. Since the loop filter is obviously strongly low pass, this means that the modulator searches candidates for the one producing the minimum baseband mean square error. Since the TNS modulator has much higher computational complexity than the SDM, it was not feasible to do a full set of 2048 simulation runs with 2^{21} sample length. The simulations are thus limited to a smaller subset of input levels. The input levels are distributed equally along the input range, with the

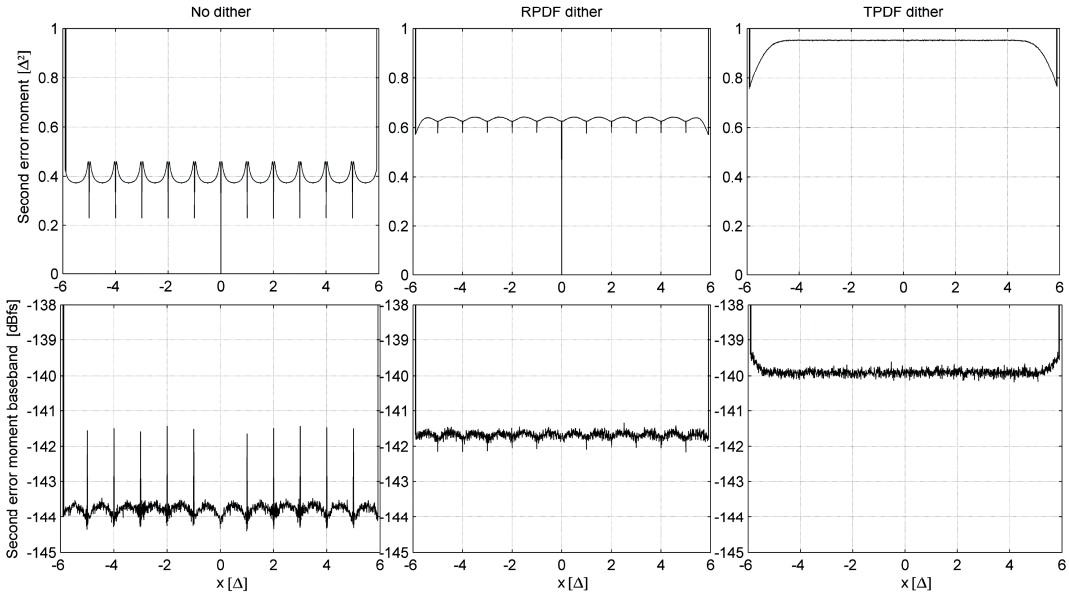


Fig. 17. Simulated noise power modulation, fifth-order 4-bit SDM.

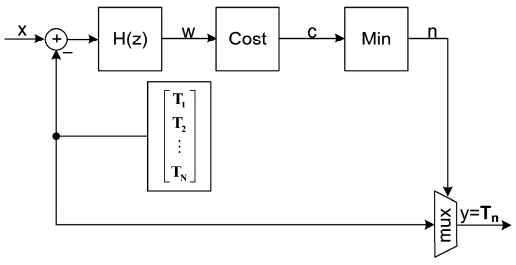


Fig. 18. General look-ahead SDM.

addition of extra runs for ±Δ/128, the level for which the regular SDM with the same NTF produced the largest in-band tones.

As can be seen from Fig. 19, the noise power modulation behavior is similar to that of the traditional SDM with the same loop filter. The total error power follows the same $(1 - x^2)$ pattern, and the baseband error increases for larger dc input levels, as also seen in the SDM case. However, the stable input range is significantly higher, with the modulator now being stable to beyond ±0.7Δ (+3 dB_{SACD}). As a consequence the dynamic range has increased by 3 dB compared to the undithered SDM, with the in-band noise power now being below −117 dBfs for a ±0.6Δ input range. Since the TNS structure is intrinsically more stable, a more aggressive NTF can be chosen for an even higher dynamic range. Furthermore the TNS modulator has no sudden increase in baseband error power for ±Δ/128 or any other input level, suggesting an absence of any in-band tones. It is also noteworthy that even beyond the stability input range the in-band noise power increases quite gently, being well below −100 dBfs even for ±0.8Δ (+4 dB_{SACD}) input.

In conclusion, the simulations suggest that the TNS modulator does not eliminate or change the basic structure of 1-bit noise power modulation. The total error power still follows the $(1 - x^2)$ pattern of the other 1-bit modulators and the in-band noise power still increases with increasing input level. However, compared to dithering it does facilitate tone-free behavior combined with better stability and a significantly higher dynamic range.

3 CONCLUSIONS

In this paper the main intention has been to supplement the ongoing discourse of dithering and SDM noise power modulation with practical and relevant simulation results for realistic modulators. Attention has been paid especially to the noise power modulation in the baseband, since this is the region where low distortion is critical.

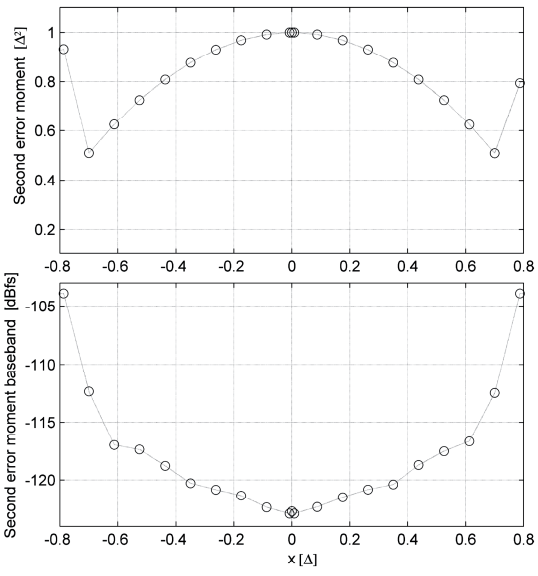


Fig. 19. Simulated noise power modulation, 1-bit TNS modulator with fourth-order trellis and fifth-order loop filter.

It has been confirmed that a multibit quantizer will need full TPDF dither to yield no noise power modulation, even if it is situated in a complex high-order sigma–delta loop. The assumption that an SDM with *N*th-order dithering is identical to an LPCM quantizer with $(2N + 1)$th-order dithering has proved erroneous. However, in very high-order SDMs, such as the fifth-order 4-bit variant, noise power modulation can be expected to be relatively small. If baseband quantization noise is already negligible, it is probably of limited or no practical interest. On the other hand in the third-order case the noise power modulation was significant. This shows that noise power modulation performance is something that should be simulated for each implementation when choosing the loop filter and dithering strategy.

In the case of 1-bit SDMs it has been confirmed through simulations that no dither will change the input dependence of the total noise power. The sole point of dithering in 1-bit SDMs should consequently be to eliminate tonal behavior; the converter will always exhibit noise power modulation. In the baseband high dither levels can actually be expected to increase noise power modulation due to increased quantizer overload. In conclusion, in the 1-bit case dithering should be used to ensure that the idle-tone performance of the converter is satisfactory while retaining sufficient dynamic range. There is no point in "over-dithering." It should also be noted that since the baseband noise power of the 1-bit modulator increases with increasing signal level, its noise power modulation is likely to be quite friendly from an audibility point of view. It has also been shown that dynamic dithering is a viable approach to achieve a higher dynamic range, less tonal behavior, and less noise power modulation. Since the 1-bit modulator operates in the overload range, the nonconstant dither power is mostly of theoretical interest, and the baseband noise power modulation performance is improved compared to constant high-level dithering. The trellis-type 1-bit modulator was found to have the same fundamental properties as the traditional SDM in terms of noise power modulation, but it offers the possibility for a further increase in dynamic range and a better stability compared to quantizer dithering.

For future studies the authors would recommend a more comprehensive investigation of the TNS approach, including optimized loop filters, different trellis orders, alternative cost functions, as well as the efficient algorithms proposed in [23]–[25]. We also feel it would be useful to show simulated or measured noise power modulation performance when publishing sigma–delta converter implementations.

4 ACKNOWLEDGMENT

This work was supported by the Norwegian Research Council under grant 162101 SPECK. The authors would like to thank Bjørnar Hernes of Nordic Semiconductor

ASA for guidance and to acknowledge the reviewers for valuable advice and feedback.

5 REFERENCES

[1] B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," Sc.D. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA (1956 June).

[2] W. R. Bennett, "Spectra of Quantized Signals," *Bell Sys. Tech. J.,* vol. 27, pp. 446–472 (1948 July).

[3] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dither; A Theoretical Survey," *J. Audio Eng. Soc.,* vol. 40, pp. 355–375 (1992 May).

[4] R. A. Wannamaker, "Dither and Noise Shaping in Audio Applications," M.Sc. Thesis, Dept. of Physics, University of Waterloo, Waterloo, ON, Canada (1990 Dec.).

[5] R. M. Gray, "Quantization Noise Spectra," *IEEE Trans. Inform. Theory,* vol. 36 (1990 Nov.).

[6] J. Reiss and M. Sandler, "Dither and Noise Modulation in Sigma–Delta Modulators," presented at the 115th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 51, pp. 1239, 1240 (2003 Dec.), convention paper 5935.

[7] J. A. S. Angus, "Achieving Effective Dither in Sigma–Delta Modulation Systems," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, p. 544 (2001 June), convention paper 5393.

[8] M. Gerzon and P. G. Craven, "Optimal Noise Shaping and Dither of Digital Signals," presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 37, p. 1072 (1989 Dec.), preprint 2822.

[9] J. Vanderkooy and S. P. Lipshitz, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, pp. 544, 545 (2001 June), convention paper 5398.

[10] S. P. Lipshitz and J. Vanderkooy, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators—Part 2," presented at the 111th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, p. 1231 (2001 Dec.), convention paper 5477.

[11] J. Vanderkooy and S. P. Lipshitz, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators—Part 3," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 50, pp. 516, 517 (2002 June), convention paper 5620.

[12] J. Vanderkooy and S. P. Lipshitz, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators—Part 4," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 52, p. 806 (2004 July/Aug.), convention paper 6093.

[13] S. P. Lipshitz and J. Vanderkooy, "Why 1-Bit Sigma–Delta Conversion Is Unsuitable for High-Quality Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, p. 544 (2001 June), convention paper 5395.

[14] S. P. Lipshitz and J. Vanderkooy, "Dither Myths and Facts," presented at the 117th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 53, p. 110 (2005 Jan./Feb.), convention paper 6279.

[15] R. A. Wannamaker, "The Theory of Dithered Quantization," Ph.D. Dissertation, Dept. for Applied Mathematics, University of Waterloo, Waterloo, ON, Canada (1997 June).

[16] R. A. Wannamaker, "Subtractive and Nonsubtractive Dithering: A Mathematical Comparison," *J. Audio Eng. Soc.,* vol. 52, pp. 1211–1227 (2004 Dec.).

[17] M. Annovazzi, V. Colonna, G. Gandolfi, F. Stefani, and A. Baschirotto, "A Low-Power 98-dB Multibit Audio DAC in a Standard 3.3-V 0.35-μm CMOS Technology," *IEEE J. Solid-State Circuits,* vol. 37, pp. 825–834 (2002 July).

[18] H. Takahashi and A. Nishio "Investigation of Practical 1-bit Delta–Sigma Conversion for Professional Audio Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, p. 544 (2001 June), convention paper 5392.

[19] S. R. Norsworth, "Dynamic Dithering of Delta–Sigma Modulators," presented at the 99th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 43, p. 1091 (1995 Dec.), preprint 4103.

[20] J. A. S. Angus, "A New Method of Applying High Levels of Dither to Sigma–Delta Modulators," presented at the 117th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 53, p. 116 (2005 Jan./Feb.), convention paper 6296.

[21] H. Kato, "Trellis Noise-Shaping Converters and 1-bit Digital Audio," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 50, p. 516 (2002 June), convention paper 5615.

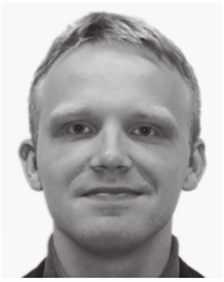
[22] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inform. Theory,* vol. IT-13, pp. 260–269 (1967).

[23] P. Harpe, D. Reefman, and E. Janssen, "Efficient Trellis-Type Sigma–Delta Modulator," presented at the 114th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 51, p. 439 (2003 May), convention paper 5845.

[24] E. Janssen and D. Reefman, "Advances in Trellis-Based SDM Structures," presented at the 115th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 51, p. 1257 (2003 Dec.), convention paper 5993.

[25] J. A. S. Angus, "Efficient Algorithms for Look-Ahead Sigma–Delta Modulators," presented at the 115th Convention of the Audio Engineering Society, *J. Audio. Eng. Soc. (Abstracts),* vol. 51, p. 1245 (2003 Dec.), convention paper 5950.

THE AUTHORS



I. Løkken

A. Vinje

T. Sæther

Ivar Løkken was born in Lillehammer, Norway, in 1979. He received a B.Sc. degree in electrical engineering from Sør-Trøndelag University College, Trondheim, Norway, in 2002 and an M.Sc degree in electrical engineering from the Norwegian University of Science and Technology, Trondheim, in 2004. He is presently with the Circuit and Systems Group at the Norwegian University of Science and Technology, working toward a Ph.D. degree on high-resolution audio digital-to-analog converters. His research interests include audio data converters and audio signal processing.

●

Anders Vinje was born in Trondheim, Norway, in 1979. He received an M.Sc. degree in electrical engineering from the Norwegian University of Science and Technology, Trondheim, Norway, in 2004. He is presently with the Circuit and Systems Group at the Norwegian University of Science and Technology, working toward a Ph.D. degree on high-speed, high-resolution analog-to-digital converters. His main research interests include analog-to-digital converters and related signal processing.

●

Trond Sæther was born in Ålesund, Norway, in 1958. He received M.Sc and Ph.D degrees in electrical engineering from the Norwegian University of Science and Technology, Trondheim, in 1981 and 1991, respectively.

From 1981 to 1983 he worked as a research scientist in the Microelectronics Group at SINTEF, in Trondheim. In 1983 he cofunded Nordic Semiconductor ASA, and he is now technical director. In 1999 July he joined the Circuits and Systems Group at the Norwegian University of Science and Technology as a part-time professor. His research interests include analog-to-digital converters and microelectromechanical systems

Appendix 3

Paper I Errata:

Errata: I.Løkken, A.Vinje, T.Sæther, ”Noise Power Modulation in Dithered and Undithered High-Order Sigma-Delta Modulators”, *Journal of the Audio Engineering Society.,* vol. 54, pp. 841–854 (2006 Sept.)

- On p. 843, in the first full paragraph, the first sentence should have read: In audio we are interested in having no harmonic distortion and no noise power modulation.
- In Section 2.3.1, p. 849, the text starting on line 14 should have read: At the same time the tonal performance is comparable to the statically dithered version. The noise power modulation and overall dynamic range are also improved significantly as compared to the statically dithered case because. . . .
- Fig.1, p.842 should have read: Midthread quantizer.
- Eq.22 should have negative sign.
- In fig.10; the x-axis range should be $\pm 0.5\Delta$ or ± 1 in absolute values, since $\Delta=2$. The same applies for figures 12-14 and 19.

Thanks to Prof. Stanley P. Lipshitz for feedback and comments and JAES chief editor Gerri Calamusa for printing parts of this errata in JAES vol.54, no.10.

Appendix 4

Paper II:

I. Løkken, A. Vinje, T. Sæther, B. Hernes: "Quantizer Nonoverload Criteria in Sigma-Delta Modulators", *IEEE Trans. Circuits and Systems Part II: Express Briefs*, vol.53, no.12, pp. 1383-1387, (2006 Dec.)

© 2006 IEEE. Reprinted, with permission, from IEEE Transactions on Circuits and Systems Part II: Express Briefs (ISSN 1549-7747)

IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, VOL. 53, NO. 12, DECEMBER 2006 1383

Quantizer Nonoverload Criteria in Sigma–Delta Modulators

Ivar Løkken, Anders Vinje, Trond Sæther, and Bjørnar Hernes

***Abstract*—A simple method to guarantee absolute stability in multibit sigma–delta modulators (SDMs) is to ensure that the quantizer never overloads. This applies to any SDM. Derivation of the requirements for nonoverload have previously been shown for different types of modulators; the sigma–delta or output-feedback modulator with rounding quantizer as well as the error-feedback modulator using truncation. Here, these nonoverload requirements will be clarified and a unified formulation is presented that is not limited with regard to modulator topology or quantizer function.**

***Index Terms*—Data conversion, delta–sigma, high order, overload, sigma–delta, stability.**

I. Introduction

If the quantizer input in a sigma–delta modulator (SDM) never overloads, the quantization error will be bound. Hence, if both the signal transfer function (STF) and the noise transfer function (NTF) are bound-input-bound-output (BIBO) stable and the input signal is bound, the SDM is guaranteed to be stable.

It should be noted that the above condition for stability is sufficient, but not necessary. It is valid for both output feedback (OF) modulators and error feedback (EF) modulators, which are shown in their most generalized form in Fig. 1. The quantizer is modelled as an additive error source e. In reality the quantization error is a deterministic function of the input, i.e., $e(x)$, but usually it is considered to be a random additive noise source for easier statistical analysis of the system. This brief does not treat any statistical properties of the quantizer error except its mean and boundaries, so this is not of any concern here. The input x is assumed to be stationary and zero mean. In [1] the nonoverload requirement is formalized with special emphasis on an Nth-order OF modulator with an NTF of $(1-z^{-1})^N$. It is proved that this modulator will have no overload if the quantizer has $B \geq N+1$ bits output and the input x is bound to less than half the quantizer range. This result only applies if an ideal zero-mean or rounding quantizer is used. By introducing quantizer offset or a truncating quantizer, the condition will not hold.

Manuscript received June 7, 2006; revised August 11, 2006. This work was supported in part by the Norwegian Research Council under Grant 162101 SPECK. This paper was recommended by Associate Editor Associate Editor G. Manganaro.

I. Løkken, A. Vinje, and T. Sæther are with the Norwegian University of Science and Technology, Department of Electronics and Telecommunications, Trondheim, NO 7491 Norway (e-mail: ivar.loekken@iet.ntnu.no; anders.vinje@iet.ntnu.no; trond.saether@iet.ntnu.no).

B. Hernes is with Nordic Semiconductor ASA, Tiller, NO 7075 Norway (e-mail: bjornar.hernes@nordicsemi.no).

Digital Object Identifier 10.1109/TCSII.2006.885967

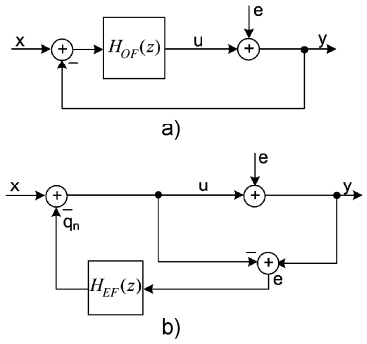


Fig. 1. OF modulator (a) and EF modulator (b).

In [2], a more informal analysis is done on an EF modulator with a truncating quantizer. It is proved that with an NTF of $(1-z^{-1})^N$, it will have no overload if $B \geq N+1$ bits output and x is bound to less than half the quantizer range. It is also shown through simulations that an OF modulator with the same NTF is not stable for the same criteria.

From this, it is tempting to conclude that the EF modulator is intrinsically more stable than the OF modulator and can tolerate larger quantization errors. However, we will show that this is not necessarily the case. We will also clarify how the quantizer function affects the nonoverload requirement and extend the formulation in [1] to include any quantizer offset.

II. Nonoverload Requirement Calculations

In this section, we will review the calculations leading to the nonoverload criteria, and their application to both OF and EF modulators. These have previously been presented in various forms in [1] as well as [4]–[7].

A. OF SDM

Looking at the OF modulator in Fig. 1(a), we can easily find that the input–output relation is given by

$$\begin{aligned} Y(z) &= \frac{H_{\mathrm{OF}}(z)}{1+H_{\mathrm{OF}}(z)}X(z) + \frac{1}{1+H_{\mathrm{OF}}(z)}E(z) \\ &= \mathrm{STF}(z)X(z) + \mathrm{NTF}(z)E(z). \end{aligned} \tag{1}$$

Here, the NTF and the STF are defined as functions of the loop filter $H_{\mathrm{OF}}(z)$. The expression for the quantizer input u can then be found as

$$\begin{aligned} U(z) &= Y(z) - E(z) \\ &= \mathrm{STF}(z)\cdot X(z) + (\mathrm{NTF}(z)-1)\cdot E(z). \end{aligned} \tag{2}$$

1057-7130/$20.00 © 2006 IEEE

Through the inverse z-transform, (2) can be transformed from the frequency domain to the time domain

$$u[n] = \sum_k stf[k] \cdot x[n-k] + \sum_k ntf[k] \cdot e[n-k] - e[n]. \quad (3)$$

By applying Schwartz' inequality, and for simplicity writing the terms as Λ-norms, the maximum quantizer input amplitude is found to be limited by

$$\begin{aligned} |u[n]| &\le \left| \sum_k stf[k] \cdot x[n-k] \right| \\ &\quad + \left| \sum_k ntf[k] \cdot e[n-k] \right| - |e[n]| \\ \|u\|_\infty &\le \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e\|_\infty - \|e\|_\infty. \end{aligned} \quad (4)$$

The maximum quantizer input amplitude occurs if there is equality in (4) and this must be smaller than the quantizer input range, R. Hence

$$|R| \ge \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e\|_\infty - \|e\|_\infty. \quad (5)$$

Equation (5) is the requirement for no quantizer overload. If the Λ_1-norms of the NTF and STF can be found and we know the maximum amplitude of the input signal x and the quantizer error e in the nonoverload case, we have a formal requirement for a bound system where no overload can occur. Hence, a modulator fulfilling (5) is guaranteed to be stable.

Since this condition is sufficient, but not necessary, many SDM converters are designed more aggressively, frequently operating in the overload range. For modulators operating in the overload range, stability analysis is very difficult and one usually relies on simulations or empirical results such as Lee's rule [3].

B. Error Feedback Modulator

The analysis of error feedback modulators is almost identical. Looking at Fig. 1(b), it is seen that the input–output relation can be expressed as

$$\begin{aligned} Y(z) &= X(z) + (1 - H_{\mathrm{EF}}(z))\, E(z) \\ &= X(z) + \mathrm{NTF}(z) E(z). \end{aligned} \quad (6)$$

Repeating the steps in (2) to (5) then results in

$$|R| \ge \|x\|_\infty + \|ntf\|_1 \cdot \|e\|_\infty - \|e\|_\infty. \quad (7)$$

Except for the absence of an STF, which is always 1 in error feedback modulators, the requirement is identical to that derived in Section II-A. This means that for an OF modulator with a unity STF, (5) equals (7), and it will behave identically to the error feedback modulator. This contradicts the conclusion found in [2]. We will assess this in the following section.

III. $(1 - z^{-1})^N$ NTF of Modulator

We will now take a closer look at the OF modulator with a $(1 - z^{-1})^N$ NTF. This is usually implemented with the loop filter consisting of a cascade of $N - 1$ delay-free integrators and one delaying integrator, as shown in Fig. 2. This is the same modulator that is analyzed in [1].

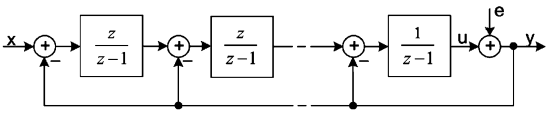


Fig. 2. $(1 - z^{-1})^N$ OF modulator.

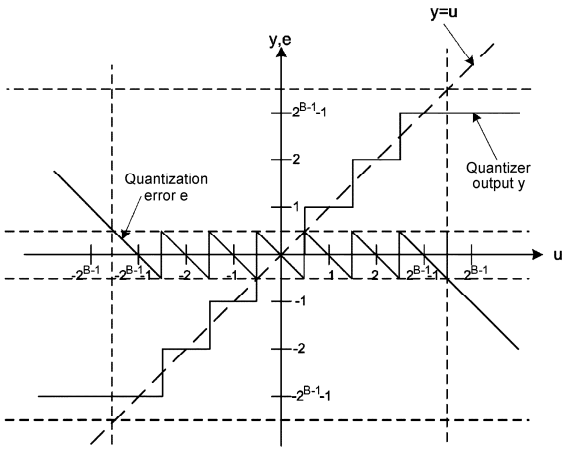


Fig. 3. B-bit symmetric $2^B - 1$-level rounding quantizer $B = 3$.

The input–output relation for this modulator is found to be

$$Y(z) = \frac{1}{z} X(z) + \frac{(z-1)^N}{z^N} E(z). \quad (8)$$

Hence the NTF and STF are given by

$$\mathrm{STF}(z) = z^{-1} \rightarrow \|stf\|_1 = 1 \quad (9)$$

$$\mathrm{NTF}(z) = (1 - z^{-1})^N \rightarrow \|ntf\|_1 = 2^N. \quad (10)$$

We furthermore assume a B-bit, $(2^B - 1)$-level symmetric rounding quantizer with $B \ge 2$, as shown in Fig. 3. The nonoverload range R is given by the input for which the quantization error is bound, which will be

$$|R| = \left(2^{B-1} - \frac{1}{2}\right). \quad (11)$$

Assuming no overload, the error is bound by

$$\|e\|_\infty = \frac{1}{2}. \quad (12)$$

We assume the input signal to occupy half the quantizer range, i.e., one bit more on the quantizer input than on the input signal. The nonoverload criterion can then be calculated as shown

$$\|x\|_\infty = 2^{B-2} \quad (13)$$

$$2^{B-1} - \frac{1}{2} \ge 2^{B-2} + 2^N \cdot \frac{1}{2} - \frac{1}{2} \rightarrow B \ge N + 1. \quad (14)$$

This is the same result found in [1]; an Nth-order cascade of integrators OF SDM where $B \geq N+1$ will not exhibit overload for input up to half the quantizer range. Also, it is the same result as we would expect from an EF modulator, as $\|\mathrm{stf}\|_1 = 1$ inserted in (5) equals (7). Indeed, in [2] these results are confirmed for the EF modulator. However, [2] also reports inferior stability for OF modulators with the same NTF.

The reason the simulation results for OF versus EF modulators in [2] differed, is the implementation of the OF modulator. In this publication, the $(1-z^{-1})^N$ OF modulator was implemented deriving the loop filter $H_{\mathrm{OF}}(z)$ directly from (1) which gives

$$\begin{aligned}\mathrm{NTF}(z) &= \frac{1}{1+H_{\mathrm{OF}}(z)} \\ &= \frac{(z-1)^N}{z^N}\end{aligned} \rightarrow H_{\mathrm{OF}}(z) = \frac{z^N-(z-1)^N}{(z-1)^N}. \tag{15}$$

When this is inserted as the loop filter in the structure from Fig. 1(a), the NTF will be the same as for the modulator in Fig. 2, but the STF will no longer be unity

$$\begin{aligned}\mathrm{STF}(z) &= \frac{H_{\mathrm{OF}}(z)}{1+H_{\mathrm{OF}}(z)} \\ &= \frac{z^N-(z-1)^N}{z^N}\end{aligned} \rightarrow \|stf\|_1 = 2^N - 1. \tag{16}$$

This gives a different nonoverload requirement than that of the distributed feedback OF modulator in Fig. 2. Looking at (5), it is clear that both the NTF and the STF affect the overload conditions in a SDM. In most literature, stability is only treated in the NTF context as the STF is assumed unity. The peak gain of the STF of (16) is much higher than unity, compromising the nonoverload range and leading to an erroneous conclusion that OF modulators are inherently less stable than EF modulators [2]. The STF in (16) has steadily increasing gain with higher frequency, peaking at $2^N - 1$ at half the sampling frequency. Thus, it can be expected to cause inferior stability for low oversampling ratio modulators.

IV. Rounding and Truncating Quantizers

The observant reader has perhaps noted another apparent discrepancy in the results found so far. If the quantizer is a truncator, then the result from (14) will definitely not hold. A truncator has a transfer characteristic as shown in Fig. 4 and hence $\|\mathrm{e}\|_\infty = 1$ and $|\mathrm{R}| = 2^{B-1}$. Inserted into (7) this gives

$$2^{B-2} + 2^N \cdot 1 - 1 \leq 2^{B-1} \rightarrow B \geq N+2. \tag{17}$$

This will be the same for both the EF modulator and the OF modulator from Fig. 2. Still, as mentioned in the introduction, the findings in [2] have proved the EF modulator with truncation to have no quantizer overload for the given conditions if $B \geq N+1$. The analysis in [2] is based on the growth in word length from each arithmetic operation in the modulator. For the EF modulator this is easily derived, since both NTF(z) and $H_{\mathrm{EF}}(z) = 1 - \mathrm{NTF}(z)$ are finite-impulse response (FIR) filters (it should be noted that for the OF modulator, having an infinite-impulse response (IIR) loop filter, the particular method in [2] can not be used).

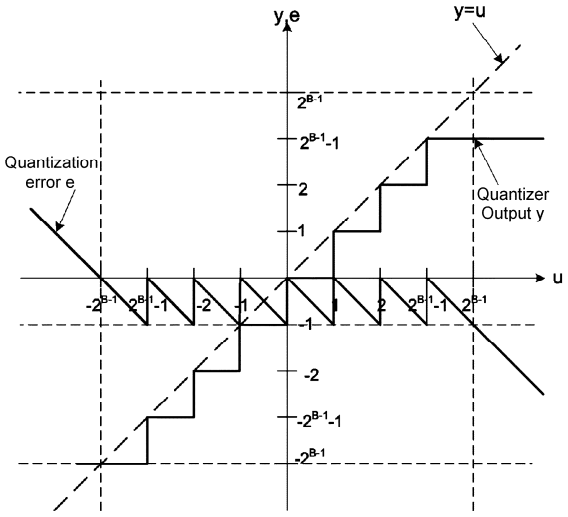


Fig. 4. B-bit 2^B-level truncating quantizer, $B = 3$.

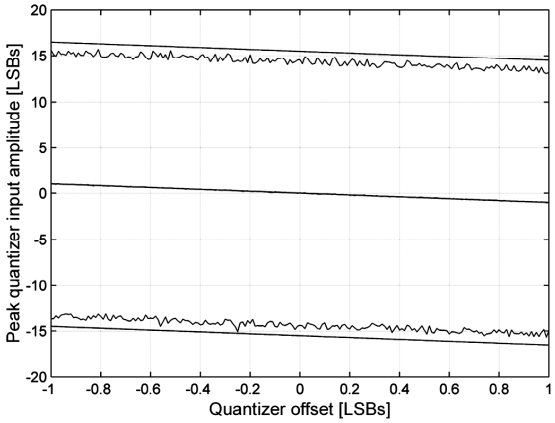


Fig. 5. Simulated peak quantizer input.

However, this discrepancy can be explained, and it turns out that both the OF modulator and the EF modulator will indeed fulfil the nonoverload requirement from (14) also with a truncating quantizer. We repeat the general formula for the peak quantizer input

$$\|u\|_\infty \leq \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e\|_\infty - \|e\|_\infty. \tag{18}$$

Given $\|stf\|_1 = 1$, this is also valid for the EF modulator. From Fig. 4, it can be seen that compared to the rounding operation, the error from the truncator has a dc offset; $\overline{e}$. We can divide e into e_{ac} and $\overline{e}$, noting that $|e|^+_{\max} = \|e_{\mathrm{ac}}\|_\infty + \overline{e}$ and $|e|^-_{\max} = \|e_{\mathrm{ac}}\|_\infty - \overline{e}$. Since e is directly visible on the quantizer input, negated, u is offset by the amount $\overline{e}$, meaning that the positive and negative maxima of u will not be equal.

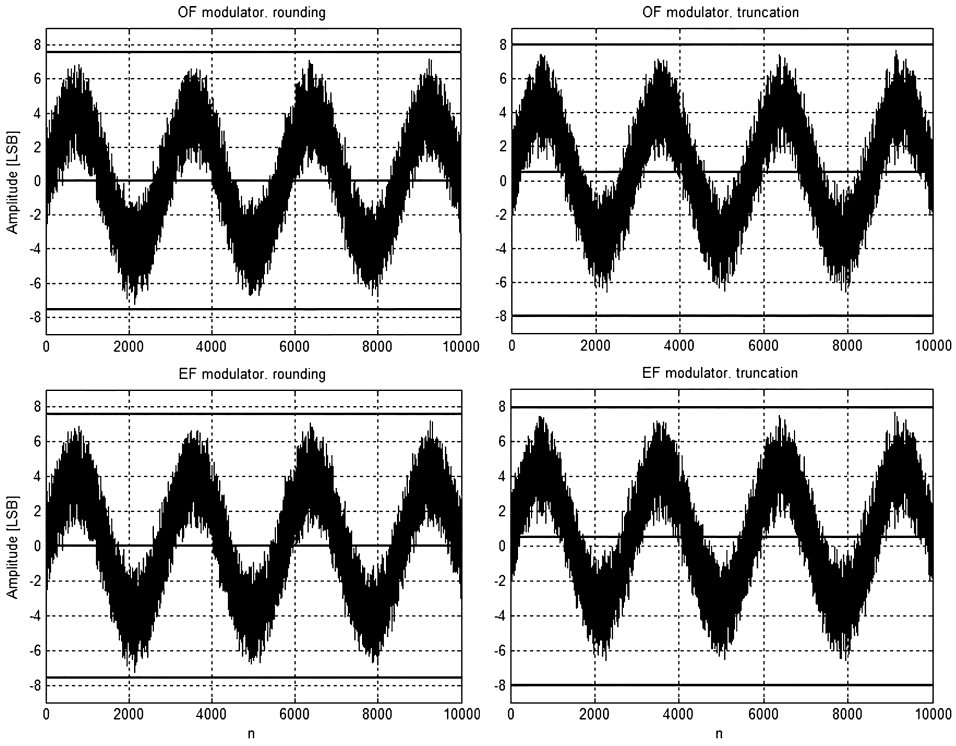


Fig. 6. Simulated quantizer input, third-order modulators with 4-bit quantization.

Also, and more importantly, the NTF is zero at dc. Hence, the contribution $\|ntf\|_1 \cdot \|e\|_\infty$ in (18) should be replaced with $\|ntf\|_1 \cdot \|e_{\rm ac}\|_\infty$. We thus end up with

$$|u|^+_{\max} \le \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_{\rm ac}\|_\infty - \|e_{\rm ac}\|_\infty - \overline{e} \quad (19a)$$

$$|u|^-_{\max} \le \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_{\rm ac}\|_\infty - \|e_{\rm ac}\|_\infty + \overline{e}. \quad (19b)$$

Here, $\|e_{\rm ac}\|_\infty$ will be given by the maximum quantizer error after it is compensated for static offset. Hence, for no overload, the requirement for the quantizer range is

$$|R|^+ \ge \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_{\rm ac}\|_\infty - \|e_{\rm ac}\|_\infty - \overline{e} \quad (20a)$$

$$|R|^- \ge \|stf\|_1 \cdot \|x\|_\infty + \|ntf\|_1 \cdot \|e_{\rm ac}\|_\infty - \|e_{\rm ac}\|_\infty + \overline{e}. \quad (20b)$$

This new inequality (20) will be valid for both rounding and truncation, or any quantizer with static offset up to ± 1 LSB (above which, it will begin to drop levels). It is also valid for OF modulators as well as EF modulators (for the latter, insert $\|stf\|_1 = 1$) and thus represents the unified requirement mentioned in the introduction.

With the B-bit truncator as shown in Fig. 4, $|R|^+ = |R|^- = 2^{B-1}$, $\|e_{\rm ac}\|_\infty = 1/2$ and $\overline{e} = -1/2$. Inserting these values in the new nonoverload requirement and applying it to the modulator in Section III results in the minimum requirement

$$2^{B-1} \ge 2^{B-2} + 2^N \cdot \frac{1}{2} \rightarrow B \ge N + 1. \quad (21)$$

With the new modified nonoverload requirement, the theory is in conformance with the results found for the truncating EF modulator in [2].

V. Simulations

The new nonoverload requirement has been evaluated by simulation on register transfer level (RTL) models in Matlab.

Fig. 5 shows $u_{\max}$, $u_{\min}$ and $\overline{u}$ as a function of the quantizer offset $\overline{e}$. The input signal x is a uniformly distributed pseudorandom sequence of length 2^{16} samples, with peak value 0.5 normalized to the quantizer input range. A new pseudorandom sequence was generated for each of the 200 iterations of the quantizer offset value. The modulator is a fourth-order, 5-bit OF modulator designed according to Fig. 2.

The theoretical limits according to (19) are shown together with the simulated quantizer peak input. As can be seen from the figure, the pseudorandom input sequence generates quantizer input close to the limits given by (19), but never exceeding them. Both the mean and peak values can be seen to adhere to (19) as a function of offset.

Fig. 6 shows the simulated input to the quantizer in an OF modulator as well as an EF modulator. In both cases, the NTF

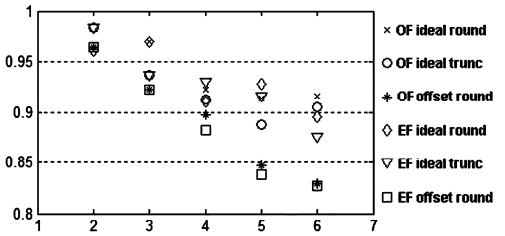


Fig. 7. Peak input normalized to nonoverload range, second- to sixth-order modulators.

is $(1 - z^{-1})^3$ and the STF is unity. The OF modulator is implemented according to Fig. 2. The input signal is a sinusoid with amplitude of half the quantizer range. The quantization functions shown are 4-bit rounding and truncation. It can be seen that the quantizer input is practically identical for the OF case and the EF case, and in all simulations within the overload limits. This supports the $B \geq N + 1$ stability requirement as valid for both types of modulator as well as both quantization functions. It can also be seen that with truncation, the quantizer input is offset and in both cases its mean value is 0.5 LSBs as expected. The simulated mean is shown as well as the limits of quantizer overload.

The results were confirmed with simulations on modulators from 1st to sixth-order as well as previously shown range of offset values. Fig. 7 shows the peak input for Nth-order $(1 - z^{-1})^N$ modulators with N ranging from 2 to 6. The number of bits B equals $N+1$ and the inputs are pseudorandom sequences with maximum amplitude determined by the nonoverload criterion from (20). Simulations are shown for ideal rounding, ideal truncation and a nonideal rounding with 0.2 LSB offset. All maximum input amplitudes are normalized to the nonoverload range. As can be seen in Fig. 7, the maximum quantizer input peak amplitude has better margins the higher the modulator order is. With higher order it will be less and less likely that a finite length input will contain exactly the sequence that excites the theoretical maximum amplitude, or amplitudes very close to it. The slight differences between OF and EF for high order modulators are also due to finite simulation lengths and have been found to be random. The new overload criterion holds in all cases.

VI. Conclusion

In this brief, the SDM nonoverload requirement has been expanded to include quantization functions with arbitrary offset. This facilitates the analysis of truncators, often used in error feedback DACs, but also quantizers with systematic offset, which can be useful for SDM ADC designers. However, it should be mentioned that in the ADC case, nonideal integrators might result in (9)–(10) not being exact as well as the NTF not being exactly zero at DC. In this case, headroom must be taken into account in the analysis, now being approximate. This will of course be implementation dependent.

We have also shown that the requirements for no overload of an EF modulator are identical to those for an OF modulator with the same NTF, provided the STF of the latter is unity. In these cases, we have also confirmed that a Nth-order modulator with a $(1 - z^{-1})^N$ NTF will be stable if the number of bits in quantizer output is $B \geq N + 1$. This applies to both the ideal rounder and the ideal truncator.

Acknowledgment

The authors would like to thank P. Kiss, Agere Systems, for contributing with valuable input and advice during the composition of this brief.

References

[1] R. Schreier and G. Temes, *Understanding Delta-Sigma Modulators*. New York: IEEE Press, 2005, ch. 4.2.2, pp. 104–107, 0-471-46585-2.

[2] P. Kiss, J. Arias, D. Li, and V. Boccuzzi, "Stable high-order delta–sigma DACs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 1, pp. 200–205, Jan. 2004.

[3] W. L. Lee, "A Novel higher-order interpolative modulator topology for high resolution oversampling A/D converters," M.Sc. thesis, MIT, Cambridge, MA, Jun. 1987.

[4] G. Bourdopoulos, A. Pnevmatikakis, V. Anastassopoulos, and T. Deliyannis, *Delta-Sigma Modulators; Modelling, Design and Applications*. London, U.K.: Imperial College Press, 2003, ch. 3.8, pp. 64–66, 1-86094-369-1.

[5] S. R. Norsworthy, R. Schreier, and G. Temes, *Delta-Sigma Data Converters; Theory, Design and Simulation*. New York: IEEE Press, 1997, ch. 3.14, pp. 130–131, 0-7803-1045-4.

[6] S. R. Norsworthy, "Optimal nonrecursive noise shaping filters for oversampling data converters, part 1: Theory," in *Proc. IEEE ISCAS'93*, May 1993, vol. 2, pp. 1353–1356.

[7] R. Schreier, "An empirical study of high-order single bit delta–sigma modulators," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 40, no. 8, pp. 461–466, Aug. 1993.

Appendix 5

Paper III:

I. Løkken, A. Vinje, T. Sæther: "Segmented Dynamic Element Matching Using Delta-Sigma Modulation", *Proc. 31st Conference of the Audio Engineering Society – New Directions is High Resolution Audio*, London UK, (2007 June).

SEGMENTED DYNAMIC ELEMENT MATCHING USING DELTA-SIGMA MODULATION

IVAR LØKKEN[1], ANDERS VINJE[2], AND TROND SÆTHER[3]

[1] *Norwegian University of Science and Technology, Trondheim, Norway*
ivar.loekken@iet.ntnu.no
[2] *Norwegian University of Science and Technology, Trondheim, Norway*
anders.vinje@iet.ntnu.no
[3] *Norwegian University of Science and Technology, Trondheim, Norway*
trond.saether@iet.ntnu.no

In multi-bit delta-sigma digital-to-analog converters (DACs), the distortion from physical element mismatch can be spectrally shaped using dynamic element matching (DEM). A problem with all DEM schemes is that the complexity increases very rapidly with the number of levels in the DAC. To reduce DEM complexity for DACs with many bits, DEM segmentation using a dedicated segmentation delta-sigma modulator (DSM) has previously been suggested. Published segmentation-DSMs have usually been first-order error feedback designs, to maximize the DEM complexity reduction and to minimize the analog overhead. In this paper high-order segmentation-DSMs will be investigated and improved solutions proposed.

INTRODUCTION

In oversampled data converters with delta-sigma modulation (DSM), one can achieve very high baseband resolution with few or only one bit. One-bit conversion avoids static nonlinearity, but in recent years multi-bit DSM has become increasingly popular for high resolution applications due to better stability, less out-of-band noise and lower sensitivity to wide band clock jitter. These advantages all increase with the number of bits in the DSM and DAC. It is therefore desirable for very high resolution audio if one can use many bits.

A major disadvantage with multi-bit DSM is that the performance will be significantly compromised by static nonlinearity from physical DAC element mismatch. To overcome this problem, digital algorithms for dynamic element matching (DEM) were developed, exploiting the redundancy of element selection in a thermometer encoded DAC. Notable contributions include Data Weighted Averaging (DWA) [1]-[3], Tree Structured DEM (TDEM) [4]-[5] and vector feedback DEM (VDEM) [6]-[7]. Implementations have been shown that feature first- and second-order high-pass spectral shaping of DAC mismatch errors.

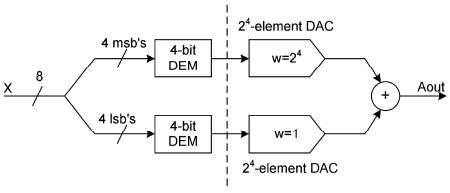


Figure 1: Segmented dynamic element matching

A problem with all known DEM algorithms is that the complexity grows exponentially with the number of bits and therefore multi-bit DSM DACs have traditionally been limited to five bits or less. This contradicts desires for high resolution. A possible solution is to segment the DAC into two weighted sub-DACs and use individual, smaller DEM blocks for each, like shown in fig.1 for an 8-bit example.

In this example, the complexity is reduced from one 8-bit DEM encoder to two 4-bit DEM encoders. However, a major problem with this approach is that even though the intra DAC mismatch is linearized with a DEM algorithm, any weighting error *between* the sub-DACs due to physical mismatch is not in any way shaped. Consequently any such weighting errors will lead to unshaped noise and signal distortion.

1 DAC SEGMENTATION IN THE LINEAR SIGNAL DOMAIN

The distortion from weighting errors is most easily analyzed by looking at the system in the linear signal domain, as shown in fig.2. Splitting the data is equivalent to introducing a truncating quantizer and the lower DAC can be viewed as a compensation DAC for the introduced quantization error.

If each sub-DAC is DEM encoded so that is has negligible distortion in the frequency band of interest, it can be considered a linear gain element equal to its weight. The coarse DAC gain is nullified by the division factor caused by right shifting the MSBs. In the signal

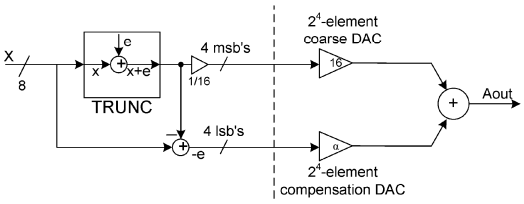


Figure 2: Segmented DEM in the linear signal domain

domain, a weighting error between the main and the compensation DAC can be seen as a non-ideal gain factor in the compensation DAC[1], replacing its ideal gain of 1 with a non-ideal factor α. Then, the output will be given by:

$$A_{out} = X + (1-\alpha)\cdot e \tag{1}$$

This means that the quantization error *e*, introduced by the 8-bit to 4-bit truncation, will leak to the analog output. Since such a coarse quantization has severe in-band error power, this leakage will cause the performance of the DAC to deteriorate significantly. The spectrum of the output distortion can be found from the error spectrum of the truncation operation, as seen from expression (1). In practice, α will typically be a random variable, Gaussian distributed around 1 and with an implementation dependent standard-deviation of e.g. 1%.

A solution to greatly reduce this distortion problem was proposed in a very high performance DAC design [8], introducing a dedicated first-order segmentation-DSM as shown in fig.3. A similar solution was also featured in a recent delta-sigma ADC [9]. From fig.3 it is seen that in this case, non-ideal weighting will lead to the output being given by:

$$A_{out} = X + (1-\alpha)\cdot e_{DSM} \tag{2}$$

Without loss of generality, a DSM can be described by its signal transfer function (STF) and noise transfer function (NTF):

$$Y(z) = STF(z)\cdot X(z) + NTF(z)\cdot E(z) \tag{3}$$

In which case the error e_{DSM} will be given by:

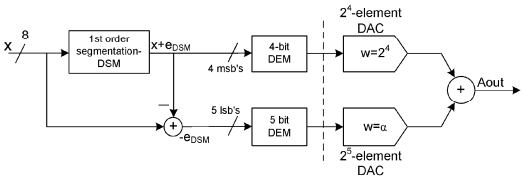


Figure 3: Segmented DEM with segmentation-DSM [8]

$$\begin{aligned} E_{DSM}(z) &= X(z) - Y(z) \\ &= (1 - STF(z))\cdot X(z) - NTF(z)\cdot E(z) \end{aligned} \tag{4}$$

Since the STF and NTF can be chosen so that e_{DSM} is a spectrally shaped error with little baseband power, leakage from weighting errors will lead to much less baseband distortion. The operation can also be repeated in multiple steps for a higher degree of segmentation [10].

It should be noted from fig.3 that the compensation DAC now has 2^5 elements instead of 2^4, since the peak error of the delta-sigma modulator, $\|e_{DSM}\|_\infty$, can grow larger than the peak truncation error $\|e\|_\infty$. Because extra bits in the compensation DAC means analog overhead as well as less reduction in DEM complexity, the segmentation-DSM shown in previous literature has typically been an error feedback design with unity STF and a first-order finite impulse response (FIR) NTF of *(z-1)*. Then it is easily seen from (4) that the peak value of e_{DSM} can never be more than twice that of *e* and the compensation DAC will need one extra bit, or twice the number of levels, like shown in fig.3.

The drawback with using a first-order segmentation-DSM is that e_{DSM} will still have quite significant baseband power and may also contain idle-tones, which in the case of weighting errors leak to the output and degrade performance. A second-order error feedback segmentation-DSM with a FIR NTF of $(z-1)^2$ will perform much better, but at the cost of increased complexity since $\|e_{DSM}\|_\infty$ can now be four times larger than $\|e\|_\infty$ and the compensation DAC will need four times more levels. Because of this, second or higher-order FIR segmentation-DSMs were discarded in [9].

2 GENERAL HIGH-ORDER SEGMENTATION DSM

As highlighted in previous work by the authors [11], the quantizer overload criteria and peak quantizer input can be found for any high-pass NTF DSM with any quantizer function, if one can find the L_1-norms of the STF and NTF impulse responses. If the segmentation-DSM is made sure to be non-overloading, the same procedure can also be used to find $\|e_{DSM}\|_\infty$. Assuming the segmentation-DSM is designed with unity STF, the first term in (4) will be zero and the worst case peak DSM error is found through:

$$\begin{aligned} &E_{DSM}(z) = -NTF(z)\cdot E(z) \\ &\Rightarrow \|e_{DSM}\|_\infty \le \|ntf\|_1 \cdot \|e\|_\infty \end{aligned} \tag{5}$$

where $\|e\|_\infty$ is the peak truncation error compensated for the truncation error mean[2], i.e. Δ/2. Rearranging (5) it is given that the increase in peak error from using a

[1] It can also be seen as a non-ideal gain factor in the main DAC. In this case, the equation will also include overall DAC gain error.

[2] The peak truncation error can be compensated for truncation error mean because the NTF is assumed to have a zero (or more) at DC, for details on this, see [11].

segmentation-DSM instead of truncation (splitting) will be:

$$\frac{\|e_{DSM}\|_\infty}{\|e\|_\infty} \le \|ntf\|_1 \quad (6)$$

This means that if the L_1-norm of the NTF is two or less, a compensation DAC implemented with one additional bit, as shown in fig.3, will always suffice. This is very restrictive and requires a first-order segmentation-DSM. However, if a relatively conservative and non-overloading NTF is chosen, a high-order segmentation-DSM with significantly better performance can be used at a much more reasonable complexity penalty than that of the previously mentioned FIR NTF of $(z\text{-}1)^2$.

As is known, the L_1-norm of the impulse response is given by the sum of all its absolute values, which for a causal and infinite impulse response (IIR) system is given by:

$$\|ntf\|_1 = \sum_{k=0}^{\infty} |ntf[k]| \quad (7)$$

Since a stable IIR converges to zero, the L_1-norm is finite and can be approximated with arbitrary accuracy by taking an arbitrarily long sample set of the sum in (7). The required number of levels for the e_{DSM} compensation DAC can then be found from:

$$n_{lev}(e_{DSM}) = \left\lceil n_{lev}(e) \cdot \|ntf\|_1 \right\rceil \quad (8)$$

This acknowledgement gives significantly larger freedom in segmentation-DSM design than just using FIR variants. Through regular DSM analysis, the spectrum of e_{DSM} can also be optimized for any desired oversampling ratio. Table I compares some example NTFs to those shown in previous literature, fig.4 shows their spectral response. All have less n_{lev} overhead than the second-order FIR and as is seen, using a conservative second-order IIR segmentation-DSM will reduce the number of levels required for the compensation DAC from 64 to 43, which is closer to first-order.

#	NTF type	NTF(z)	$\|ntf\|_1$	n_{lev}
-	Trunc	1	1	16
1	FIR1 [8]	$(z-1)$	2	32
2	FIR2 [9]	$(z-1)^2$	4	64
3	IIR2	$\frac{(z-1)^2}{z^2-1.225z+0.4415}$	2.63	43
4	IIR3	$\frac{(z-1)^3}{(z-0.6694)(z^2-1.531z+0.6639)}$	3.26	53
5	IIR4	$\frac{(z^2-1.999z+1)(z^2-1.993z+1)}{(z^2-1.49z+0.563)(z^2-1.7z+0.7861)}$	3.71	60

Table 1: Example segmentation-DSMs and properties.

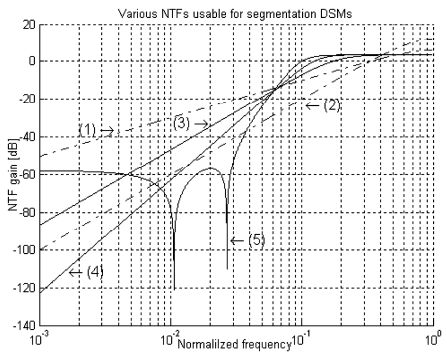


Figure 4: Spectral properties of NTFs from table 1.

The reduction in noise suppression is seen from fig.4 to be modest and also the second-order will still have major advantages in terms of tonal behaviour in e_{DSM}.

Using the non-overload method, the segmentation-DSM can be chosen with almost the same degree of freedom as the main DSM. The possibilities are many, but it should be noted that since implemented algorithms for DEM encoding of the sub-DACs are currently limited to second-order shaping, exploiting higher-order segmentation-DSMs is in this context of more theoretical than practical interest. The approach however still significantly reduces the analog overhead caused by going from a first- to second-order architecture.

3 SIMULATIONS

Since the non-overloading second-order IIR segmentation-DSM is the most realistically desirable alternative to the first-order that's been used in most previous publications, these two have been investigated further with functional-level simulations. To evaluate performance, they were first simulated with ideal sub-DACs, introducing a constant weighting error for the fine DAC of 3%, the 3σ-value for a 1% standard deviation. In all simulations the input to the DEM system was generated using an 8-bit modulator with $(z\text{-}1)^3$ NTF and the segmentation-DSM was a 4-bit unity STF modulator as shown in fig.5.

Figure 6 and 7 show the simulated spectra with a large-scale -6dBfs and a small-scale -60dBfs input signal. As can be seen in the spectra for the large-scale

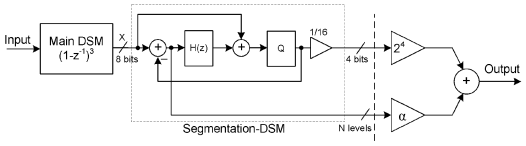


Figure 5: General simulation model with unity STF segmentation-DSM and linear sub-DACs

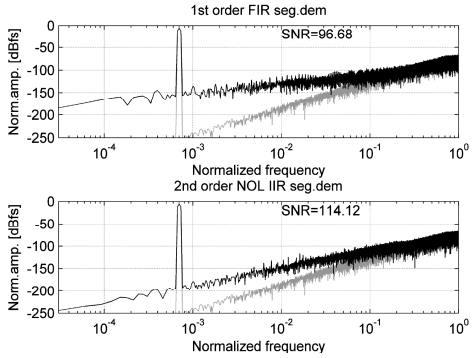


Figure 6: Simulated spectra, linear sub-DACs with 3% weighting error, -6dB input.

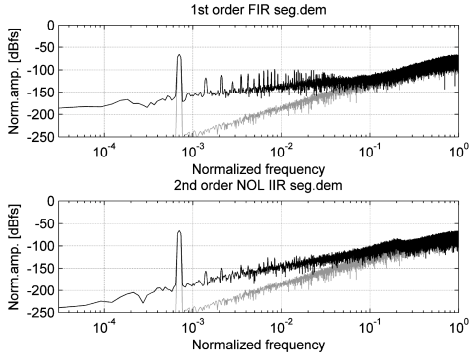


Figure 7: Simulated spectra, linear sub-DACs with 3% weighting error, -60dB input.

signal, the output is dominated by the main-DSM quantization error down to around fs/20, where it goes from third to first-order slope, suggesting it becomes dominated by leakage of the segmentation e_{DSM}. As expected it goes from third to second-order for the improved segmentation-DSM. For small level input, it is clearly seen that the leaked e_{DSM} from the first-order segmentation-DSM suffers greatly from tonal behaviour, while the second-order performs much better.

This clearly shows what advantage in low-level behaviour that can be achieved by moving to second-order, at the cost of 43 levels instead of 32 levels in the compensation DAC.

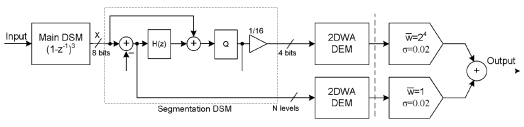


Figure 8: General simulation model with unity STF segmentation-DSM and sub-DACs with mismatch

For the next set of simulations, a more realistic model was made. Instead of modelling the sub-DACs as linear gain-elements, they were modelled as non-ideal unit element DACs, with expectance value equal to their weight and a mismatch standard deviation of 2% of the compensation DAC LSB level. This model is shown in fig.8. The same two segmentation-DSMs were used as for the previous simulations where the sub-DACs were linear gain elements. Since second-order DEM is currently the state of the art in high resolution multi-bit DSM design, each DAC was modelled utilizing a general second order DWA algorithm as described in [3]. The uppermost spectral plots in fig.9 and fig.10 shows how this algorithm performs without segmentation, that is when using a single 256-level DEM and mismatch-DAC.

As can be seen, the output spectrum again follows the third order spectrum of the main DSM down to around fs/20, below which the error is dominated by mismatch noise. The spectrum is seen to follow a clear second-order slope from there towards DC. The middle subplot shows where the DAC is segmented into a 4-bit main DAC and a 5-bit compensation DAC using a first-

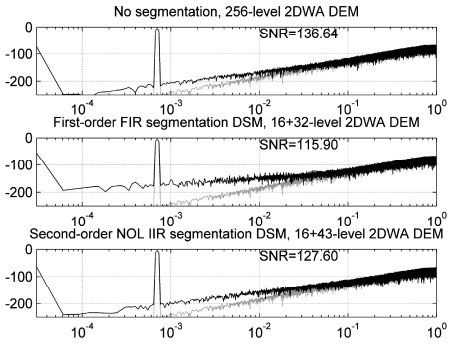


Figure 9: Simulation spectrum, segmented mismatch DAC using second order DEM, -6dB input.

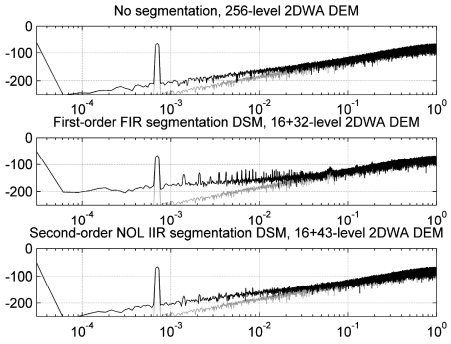


Figure 10:.Simulation spectrum, segmented mismatch DAC using second order DEM, -60dB input.

order segmentation-DSM with *(z-1)* NTF. As can be seen, the error spectrum now transits from third-order to first-order slope for low frequencies, indicating that the leakage of segmentation-DSM error dominates the total mismatch error. These findings indicate that *the segmentation-DSM in practical applications should be of the same order as the DEM used in the sub-DACs*. It is also seen in fig.10 that the output spectrum does show some tonal behaviour. With the second-order segmentation-DSM, the output spectrum is cleaner and has a second-order slope even for low-frequency errors. This shows that significantly better performance can be achieved by using 16+43-level sub-DACs and a conservative second-order segmentation-DSM, compared to 16+32-level sub-DACs and a first-order segmentation-DSM.

4 CONCLUSION

In this paper an analysis of DEM segmentation in the linear signal domain has been shown. The output distortion from analog weighting errors between the sub DACs can be directly found from the truncation error of the segmentation operation. This error can be spectrally shaped by including a separate segmentation-DSM, as has proved successful in previously publicized designs.

This paper furthermore shows that the same freedom of design is available for the segmentation-DSM as for the main DSM. Using a conservative non-overloading IIR NTF will give much better spectral shaping than a first order FIR NTF, and at a much lower cost than a second order FIR NTF.

Simulations also show that the segmentation-DSM should have the same order of shaping as the DEM algorithm used in the sub-DACs, if not it will be the Achilles' heel. A conservative second-order IIR NTF will together with second-order DEM in the sub-DACs provide a lot better performance than if *either* is first-order, at a reasonable complexity penalty.

It is likely that the development of data converters for very high resolution audio will tend towards using more bits in the DAC and main DSM. Segmented DEM facilitates this with reasonable complexity. With second-order mismatch shaping throughout, mismatch can be significant without deteriorating the end performance.

REFERENCES

[1] R.T. Baird, T.S. Fiez, "*Linearity enhancement of multi-bit Δ-Σ A/D and D/A converters using data weighted averaging*", IEEE Transactions on Circuits and Systems II, vol.42, no.12, pp.753-762, Dec.1995.

[2] Xue-Mei Gong, "*An Efficient Second-Order Dynamic Element Matching Technique for a 120-dB Multi-Bit Delta-Sigma DAC*", AES Convention Paper 5124, 108th convention of the Audio Engineering Society, Feb.2000

[3] R.K. Henderson and O. Nys, "*Dynamic element matching techniques with arbitrary noise shaping function*", in Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS), pp.293-296, May 1996.

[4] I. Galton, "*Spectral shaping of circuit errors in digital-to-analog converters*", IEEE Transactions on Circuits and Systems II, vol.44, no.10, pp.808-817, Nov.1997.

[5] J. Welz, I. Galton, E. Fogleman, "*Simplified logic for first-order and second-order mismatch-shaping digital-to-analog converters*", IEEE Transactions on Circuits and Systems II, vol.48, no.11, pp.1014-1028, Nov.2001.

[6] R. Schreier and B. Zhang, "*Noise-shaped multi-bit D/A converter employing unit elements*", Electron. Letters, vol.31, no.20, pp.1712--1713, Sept.1995.

[7] H. Lin, J. da Silva, B. Zhang, R. Schreier, "*Multi-Bit DAC with Noise-Shaped Element Mismatch*", in Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS), pp.235-238, May 1996.

[8] R. Adams, K. Nguyen, K. Sweetland, "*A 112-dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling*", AES Convention Paper 4774, 105th Convention of the Audio Engineering Society, Sept.1998.

[9] Y. Cheng, C. Petrie, B. Nordick, D. Comer, **"***Multibit Delta-Sigma Modulator With Two-Step Quantization and Segmented DAC*", IEEE Transactions on Circuits and Systems II, vol.53, no.9, pp.848–852, Sept.2006

[10] J. Stccnsgaard, "*High-resolution mismatch-shaping digital-to-analog converters*", The 2001 IEEE International Symposium on Circuits and Systems, ISCAS 2001. vol.1, Page(s):516–519, 6-9 May 2001

[11] I. Løkken, A. Vinje, T. Sæther, B. Hernes, "*Quantizer Nonoverload Criteria in Sigma-Delta Modulators*", IEEE Transactions on Circuits and Systems Part II, vol.53, no.12, pp. 1383-1387, Dec. 2006.

Appendix 6

Paper IV:

I. Løkken, A. Vinje, T. Sæther, B. Hernes: "Error Estimation in Delta-Sigma DA-Converters", *Submitted to Analog Integrated Circuits and Signal Processing*

Error Estimation in Delta-Sigma DA Converters

Ivar Løkken[1], Anders Vinje[2], Bjørnar Hernes[3] and Trond Sæther[4]

[1,2,4] Norwegian University of Science and Technology, Trondheim, Norway
ivar.loekken@iet.ntnu.no, anders.vinje@iet.ntnu.no, trond.saether@iet.ntnu.no

[3] Arctic Silicon Devices, Trondheim Norway
bjornar.hernes@arcticsilicon.com

ABSTRACT

High-resolution and very high resolution data conversion is dominated by the use of delta-sigma modulating converters. Oversampling and noise-shaping is employed to enable a coarsely quantized conversion with high effective resolution. The time-domain output waveform from a delta-sigma modulator is often impossible to predict analytically, therefore modulator design is largely based on high level digital simulations and rule-of-thumb estimation. However the output waveform also largely determines the distortion caused by analog error sources in the converter. Therefore optimization of the modulator with regards to digital quantization noise might not yield an optimal design when analog errors are included. This paper extends familiar approximation methods and estimates to include analog error sources, with the objective of providing more global rule-of-thumb optimization.

1. INTRODUCTION

In high-resolution data converters it is desirable to move as much circuitry as possible from the analog to the digital domain. In accordance with Moore's Law [1] the digital circuit density has increased exponentially with time, enabling digital signal processing (DSP) to be realized with extremely high precision. To implement analog voltage or current elements with high accuracy is on the other hand still very difficult, and neither it nor analog filtering has benefited from Moore's law. Thus the combination of oversampling and noise-shaping has become very desirable. Noise-shaping or Delta-Sigma Modulation (DSM) was first proposed by Inose and Yasuda [2]-[3]. Not much later research began into its use for oversampled data conversion [4]-[5].

The commercial breakthrough of DSM came with the introduction of the Compact Disc (CD) format [6], bringing audio into the digital world. Audio signals have very low bandwidth and extremely high resolution making them ideal for oversampled DSM. Early audio converters used high oversampling ratios (OSR) to replace complex analog antialias and reconstruction filters with digital decimation and reconstruction filters, and also to enable 1-bit noise-shaped quantization making the converter immune to static nonlinearity [7]. However the 1-bit DSM is highly sensitive to timing related errors and the force majeur in high-res today is the combination of oversampling and DSM quantizers with a more than one bit. Due to the aggressive scaling of feature sizes and supply voltages in modern CMOS, the switched capacitor approach is also increasingly receding in favour of topologies based on current-mode.

Different error sources in a data converter may put contradictory constraints on DSM design parameters, meaning that optimizing the DSM for one error effect may have adverse effects on others. This paper investigates distortion mechanisms in continuous-time DSM DACs representative for modern current-mode implementation, with emphasis on audio-range. High-level Matlab error models are used to derive estimates for how the DSM affects DAC distortion. The error estimates provide a platform to rapidly optimize DSM parameters for global DAC performance under given set of physical design conditions. The estimates could also be useful for continuous-time DSM ADCs, which are often limited in performance by their feedback DAC.

2. DELTA-SIGMA MODULATION AND SQNR ESTIMATION REVIEW

High-res data or audio is usually stored in Nyquist-rate Linear Pulse Code Modulation (LPCM) form. A DSM DAC therefore consists of an oversampler, a DSM re-quantizer (REQ) and the DAC itself. DSM theory is a highly developed subject; in this paper a review and references are provided for completeness.

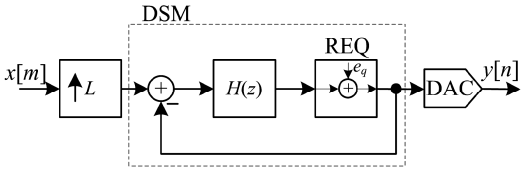


Figure 1: Delta-Sigma DAC system

A basic DSM DAC system looks like fig.1. This paper does not deal with the oversampling filter and assumes it to be ideal. In other words the DSM input is $x[n]$; the system input $x[m]$ at an L times higher sampling rate. Assuming for now the DAC is also ideal and the output is determined solely by the DSM REQ, we have:

$$\begin{aligned} Y(z) &= \frac{H(z)}{1+H(z)}X(z) + \frac{1}{1+H(z)}E_q(z) \\ &= H_{STF}(z)\cdot X(z) + H_{NTF}(z)\cdot E_q(z) \end{aligned} \quad (1)$$

STF and NTF are the DSM signal transfer function and noise transfer function respectively. If $H(z)$ is a cascade of integrators with high in-band gain the in-band STF is unity and the in-band NTF is close to zero. The more integrations, the better the DSM REQ performs. This is usually referred to as the modulator order.

The quantization error e_q can as known be approximated linearly with Bennett's noise theorem [8], stating that with unity quantization step the quantization error can be approximated as an additive white noise source with power 1/12 for random-like input signals. This approximation is widely used regardless of input signal to enable simple estimation of the signal-to-quantization noise ratio (SQNR), though more accurate models exist [9]-[11]. The power spectral density (PSD) of the DSM REQ error then becomes:

$$S_{e_{dsm}}(\omega) \approx \frac{1}{12}\cdot\frac{1}{2\pi}\left|H_{NTF}(\omega)\right|^2 \quad (2)$$

The angular frequency is $\omega=2\pi f/f_s$, where f_s is the oversampled sampling frequency $f_{s\_in}\cdot L$. If the REQ has a total of M levels, the power of a sinusoid input can be expressed as:

$$\sigma_x^2 = \frac{(k\cdot M)^2}{8} \quad (3)$$

In a normal REQ maximum k is unity (0dBfs), while in a DSM REQ it is typically between unity and 0.5 (-6dBfs) [12]. The estimated SQNR becomes:

$$SQNR \approx 10\cdot\log_{10}\left(\frac{(k\cdot M)^2}{\frac{1}{3\pi}\int_{\omega=-\frac{\pi}{L}}^{\frac{\pi}{L}}\left|H_{NTF}(\omega)\right|^2}\right) \quad (4)$$

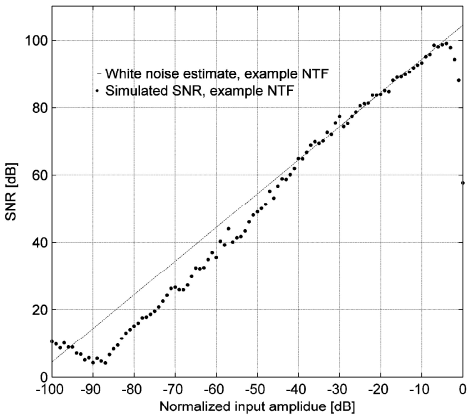


Figure 2: Simulated REQ SNR and white noise REQ SQNR estimate

In reality the quantization error e_q is not an additive white noise-source but a nonlinear function of the input. In a DSM this leads to the error PSD containing idle-tones and some distortion [13]-[14], and there may also be noise power modulation [15]. This means that the additive noise estimate is not entirely accurate and it is preferable if e_{dsm} is negligible. Figure 2 shows how the simulated SNR deviates from the predicted SQNR as a function of the input. The example is a second-order 1-bit DSM REQ with L=64. The severity of tonal behaviour and noise power modulation is reduced when the modulator order or number of bits is increased. It has historically not been trivial to achieve negligible quantization noise with a 1-bit DSM REQ. Dithering can be used [10] but at the cost of reduced maximum k. Other options include multi-stage noise shaping (MASH) or Trellis noise shaping. With multi-level quantization, it is trivial to achieve negligible in-band ε_{dsm} and such techniques are normally not necessary.

Even if the loop filter is bound-input-bound-output (BIBO) stable the quantizer introduces a nonlinearity inside the feedback loop that may cause the DSM to oscillate if k is increased beyond a certain level k<1. To obtain a reasonable stable range for k, the integrators in $H(z)$ must be damped, also damping the NTF and thus reducing the SQNR. The most used approach for this is Lee's Rule [18] stating stability will be maintained if:

$$\left\|NTF(\omega)\right\|_\infty \le 1.5 \leftrightarrow k_{\max} \le 0.5 \quad (5)$$

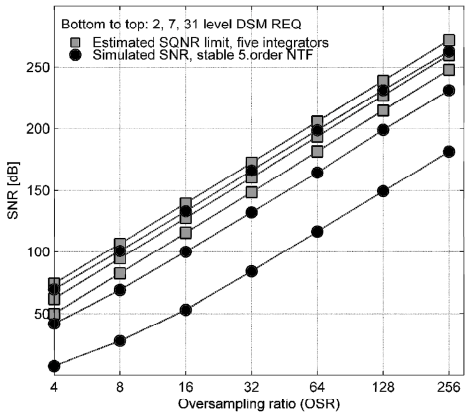


Figure 3: Example of SNR limitation due to NTF damping

Lee's rule is aimed at 1-bit modulators. Increasing M to beyond two levels means that either the NTF aggressiveness $\|NTF(\omega)\|_\infty$ or maximum k can be increased. Figure 3 shows how NTF damping reduces the SNR for a 1-bit, 3-bit and 5-bit DSM, compared to the theoretical SQNR with undamped integrators.

Despite lacking a solid mathematical basis Lee's rule is the overwhelmingly most used approach in practical DSM design. For multi-bit modulators one can increase the aggressiveness and/or peak swing gradually while simulating for stability. Rigorous mathematical stability analysis does not exist for high-order DSMs as of today, but it is an ongoing field of research in theoretical mathematics and non-linear dynamics. A summary of the current art in stability analysis is found in [19].

In an audio context, achieving state-of-the-art resolution (peak SQNR >120dB) with a 1-bit DSM REQ requires an OSR of at least 128. With e.g. a 5-bit DSM REQ it can be achieved at a much lower OSR of 16. Delta-sigma is also gaining popularity in wide-band applications that can only afford an OSR of perhaps 4-8. Then multi-bit quantization is completely necessary.

3. DAC MISMATCH, DEM AND SMNR ESTIMATION

The main drawback with multi-level quantization, historically preventing its use in very high resolution conversion is DAC element mismatch, causing a non-linear transfer function. Research into dynamic element matching (DEM) algorithms led to breakthroughs in mismatch error shaping and made multi-bit DSM very popular for both audio and higher bandwidth lower OSR high-res converters. Since mismatch is static it can be simulated digitally by modelling the DAC as a discrete amplitude mapping, preceded by a DEM algorithm. Simplified estimates are shown to predict achievable performance.

3.1. Mismatch error modelling

At the DAC output the quantization steps are not equal because of physical mismatch between DAC elements realizing their analog counterparts [20]. In a 1-bit DAC there is only a single element switching between the two REQ output levels and no mismatch will occur. In a multi-bit DAC mismatch leads to a non-linear DAC transfer function. To minimize analog mismatch the DAC is usually thermometer encoded, with $M=2^B$ (or $M=2^B-1$) equally weighted elements realizing a B-bit DAC. Thermometer encoding of y is defined by $\mathbf{t}(y)=t_0t_1\ldots\ t_{M-1}$ where $t_{0\ldots y}=1$ and $t_{y\ldots M-1}=0$.

The analog output of a DAC can be described as an amplitude mapping:

$$y_a(y)=\sum_{i=0}^{y-1} w_i \qquad (6)$$

The values w_i are element weights, ideally exactly equal to one when normalized, but in real life deviating because of mismatch. Nonlinearity is often specified by the integral nonlinearity function (INL); the deviation for value y from a straight line between zero and M-1:

$$INL(y)=y_a(y)-y\cdot\overline{w} = \sum_{i=0}^{y-1} w_i - y\cdot\overline{w} \quad , \ \overline{w}=\frac{1}{M}\sum_{i=0}^{M-1} w_i \qquad (7)$$

If the weighting error is 1% for any given element, it contributes 0.01 of an LSB to the INL. Since matching of unit current sources and capacitors in CMOS is typically limited to 0.1%-1%, the DAC can not achieve audio grade linearity.

3.2. DEM and SMNR estimation

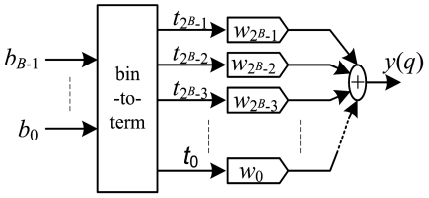


Figure 4: Generalized schematic of thermometer encoded DAC

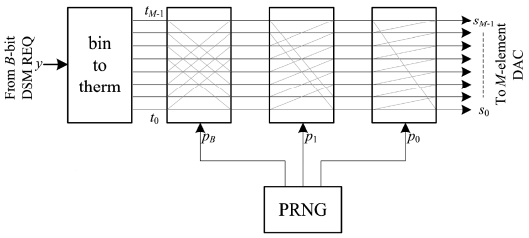


Figure 5: DEM randomization network, B=3 bit

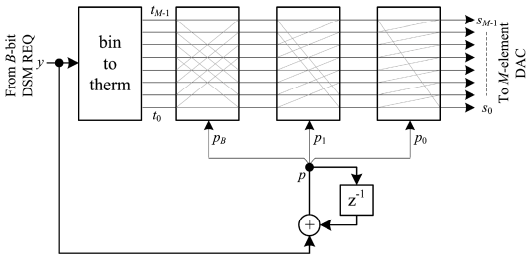


Figure 6: DWA rotation network, B=3 bit

The term DEM was introduced when Van De Plassche in 1976 [21] showed a DAC using redundant switching to improve the mismatch. In 1989 Carley published an implementation of a DEM randomizer that turned mismatch non-linearity into noise [22]. An example of a DEM randomizer for a 3-bit DAC is shown in fig.5.

The randomizer selects y random weights for sample $y[n]$ so one can't find an expression for its sample error. It can however be predicted that if element weights are random with unity expectance value and variance σ_w^2, the mismatch error expectance value is $E\{e_{mis}\}=0$ and that the mismatch error power is:

$$\begin{aligned}\sigma_{e_{mis}}^2(y) &= E\left\{\left(\sum_{i=0}^{y-1} w_i - y\cdot\overline{w}\right)^2\right\} \\ &= y\cdot\left(1-\frac{y}{M}\right)\cdot\sigma_w^2\end{aligned} \tag{8}$$

The maximum error power occurs for $y=M/2$ and is:

$$\sigma_{e_{mis}}^2 = \frac{M\cdot\sigma_w^2}{4} \tag{9}$$

Since the elements are assumed to be Gaussian random variables, the error has a white spectrum. The Wiener-Khinchin theorem can be used to find the mismatch error's PSD:

$$S_{e_{mis}}(\omega) = \frac{M\cdot\sigma_w^2}{8\pi} \tag{10}$$

This is *worst case* with randomization. Although it turns nonlinearity into more benign white noise, it does not decrease in-band error power by more than a factor L for L times OSR, making very high resolution unobtainable. This changed when element rotation algorithms were published in the 90s, data weighted averaging (DWA) [24] proving particularly successful.

With DWA the concept is for all elements to contribute equally over time, thus cancelling the integrated mismatch error. DWA is shown conceptually in fig.6.

To understand its shaping ability, imagine a selection vector **s** controlling the DAC element usage. The corresponding DAC error is then:

$$e_{mis}(\mathbf{s}) = \mathbf{s}\cdot(\mathbf{w}-\overline{w}) \tag{11}$$

The vector **w** is the element weight vector and the $\cdot$ operator is the vector dot product, meaning it is just a vector notation of the INL equation (7). To ease notation another vector **u** of length M is defined so that:

$$\mathbf{u}(a) \overset{def}{\rightarrow} u_i = \begin{cases} 1 & , 0 \le i \le a-1 \\ 0 & , a \le i \le M-1 \end{cases} \tag{12}$$

In ordinary thermometer encoding, the y lowest elements are used, so the element selection vector is:

$$\mathbf{s} = \mathbf{u}(y) \tag{13}$$

It follows that the DAC error for any given sample n is:

$$\begin{aligned} e_{mis}[n] &= \mathbf{s}[n]\cdot(\mathbf{w}-\overline{w}) \\ &= \mathbf{u}(y[n])\cdot(\mathbf{w}-\overline{w}) \\ &= \sum_{i=0}^{y[n]} w_i - y[n]\cdot\overline{w} \\ &= INL(y[n]) \end{aligned} \tag{14}$$

This means that with thermometer encoding the DAC INL directly translates to output distortion. In a DWA encoder, the element selection is rotated by updating the starting point with a rotation pointer p:

$$p[n] = (p[n-1] + y[n]) \bmod M \tag{15}$$

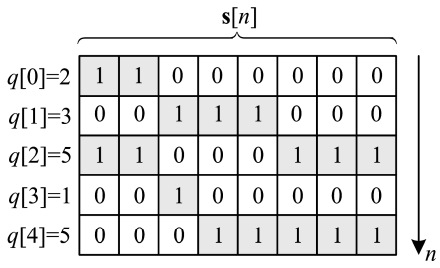


Figure 7: Element selection with DWA element rotation

The element selection vector can now be described as a function of the **u** vector as follows:

$$\mathbf{s}[n]=\begin{cases}\mathbf{u}(p[n])-\mathbf{u}(p[n-1]) & ,p[n]\geq p[n-1]\\ \mathbf{u}(M)+\mathbf{u}(p[n])-\mathbf{u}(p[n-1]) & ,p[n]<p[n-1]\end{cases} \quad (16)$$

The second case is when the element selection has rotated from the modulo pointer wrapping around *M*. Using the same procedure as for (14) the DAC error is easily found to be:

$$e_{mis}[n]=INL(p[n])-INL(p[n-1]) \quad (17)$$

Since $INL(M)=0$ (17) is true for both cases in (16). This means that the output distortion is a first order noise shaped function:

$$E_{mis}(z)=(1-z^{-1})\cdot INL(P(z)) \quad (18)$$

To derive the output spectrum analytically would require exact knowledge of the pointer's PSD. Since the pointer is directed by modulo integration it is not trivial to find. However, assuming that the input is random, $p[n]$ will also be a white random process and one can use white noise approximation for mismatch estimation just like when estimating the SQNR. For small scale input, y will be close to $M/2$ so the worst case estimate for randomization can be used to approximate the PSD of $INL(P(z))$. The mismatch error PSD is then:

$$S_{e_{mis}}(\omega)\approx\frac{M\cdot\sigma_w^2}{8\pi}\cdot\left|1-e^{i\omega}\right|^2 \approx\frac{M\cdot\sigma_w^2}{8\pi}\cdot\left|H_{DEM}(\omega)\right|^2 \quad (19)$$

The signal-to-mismatch noise ratio (SMNR) for a DWA encoded DAC can from this be estimated as:

$$SMNR\approx 10\cdot\log_{10}\left(\frac{(k\cdot M)^2}{\frac{M\cdot\sigma_w^2}{\pi}\int_{\omega=-\frac{\pi}{L}}^{\frac{\pi}{L}}\left|H_{DEM}(\omega)\right|^2}\right) \quad (20)$$

Just like the SQNR estimate, the SMNR estimate is based on the assumption of a random input signal, leading to some inaccuracy. The most serious problem it doesn't reveal is tonality: Imagine the input is a DC-signal; then $p[n]$ is clearly seen from (15) to be periodic, meaning any weighting error w_i will appear periodically and produce tones. Since the input to the DEM is typically the output from a DSM REQ, which has a strong noise component, tones are not as severe as for a first-order normal DSM, but they may be unacceptable. Dithering techniques to remove tones at the expense of shaping efficiency have been proposed [25].

Theoretical extension to second order $(1-z^{-1})^2$ DWA (2DWA) [26] is possible if the selection vector is:

$$\mathbf{s}[n]=c\cdot\mathbf{u}(M)+\mathbf{u}(p[n])-2\cdot\mathbf{u}(p[n-1])+\mathbf{u}(p[n-2]) \quad (21)$$

Then to ensure signal conservation, i.e. $\Sigma\mathbf{s}[n]=y[n]$ for all n, the pointer has to be:

$$p[n]=(2\cdot p[n-1]-p[n-2]+y[n])\bmod M \quad (22)$$

An example of a 2DWA selection vector is shown in fig.8. The problem with this is that elements take other values than just 0 and 1. This means implementation mandates DAC elements running at several times the sampling rate or being multi-level. The latter won't work since multi-level elements will themselves be non-linear, the former requires much power. It is therefore not recommended to implement 2DWA directly.

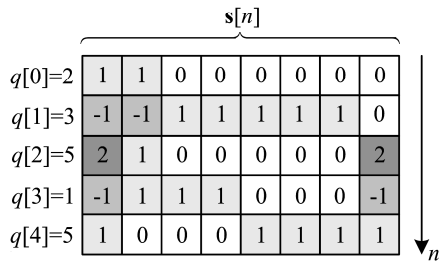


Figure 8: Element selection with 2DWA

A solution has been proposed in the form of restricted second order DWA (R2DWA) [27]. In R2DWA, an *intermediate* element selection vector is defined based on the second order noise shaping equation, and the algorithm then switches the y elements having the smallest intermediate values to "1" and the rest to "0". In other words, it uses a second order equation to determine which elements have least accumulated usage and selects them next. Since it effectively is second-order shaping with a limiter, it loses some shaping ability compared to 2DWA

3.3. SMNR estimate verification

The estimate for SMNR was verified with functional implementations of randomization, DWA, 2DWA and R2DWA algorithms.

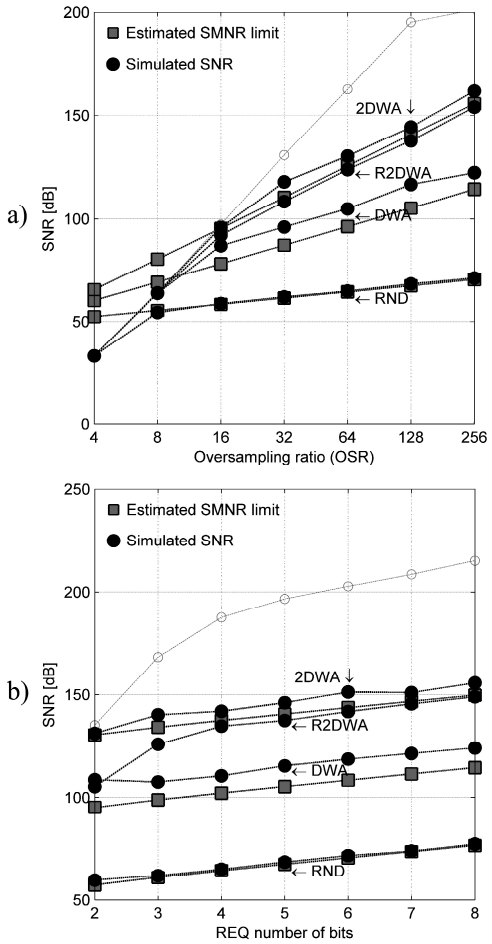


Figure 9: SMNR estimate vs. simulated SNR vs. L (a) and vs. B (b)

The simulated SNR was compared to SMNR estimates with $H_{DEM}(\omega)$ of 1, $(1-e^{i\omega})$ and $(1-e^{i\omega})^2$. The input signal was a large-scale sinusoid which, at least for the rotation algorithms, is expected to perform somewhat better than the worst-case estimate.

Selected results are shown in fig.9. The element weights are made up of a vector of Gaussian random variables with unity expectance value and 1% RMS mismatch. The a) figure has a fixed resolution of five bits and the b) figure a fixed OSR of 128. The input is a sinewave at stable full scale and $f_{s\_in}/4$. For low OSR the SNR is limited by in-band quantization noise, shown in the greyed out traces. For high OSR the mismatch noise dominates. Noise estimates are very good for randomized DEM, for first-order DWA there is some inaccuracy in the white noise approximation. The seemingly good performance may therefore be misleading since the error may be concentrated in tones. The spectrum should be simulated and assessed, but the estimate provides a good starting point to predict necessary OSR and physical mismatch σ_w for a given resolution. With second order shaping the pointer is more white and the approximation thus better. R2DWA is about 10dB inferior to ideal 2DWA and less effective with few bits since it then saturates more easily. With 1% RMS mismatch, state of the art audio resolution requires an OSR of 128 or higher with first order DEM and in the region 32-64 with second-order DEM.

It should be noted that while the SQNR increases with 6dB per bit, the SMNR increases by 3dB per bit, also apparent from (20). This assumes constant physical mismatch σ_w normalized to the LSB level. When increasing the number of bits from B to B+1 the *physical* LSB level is usually halved if the target SNR is constant. Then, as apparent from [20], the normalized mismatch variance σ_w^2 doubles, meaning that the SMNR stays the same.

The reader should also be aware that even though DWA is the most intuitive approach to understand mismatch-shaping DEM, other solutions exist that are more hardware efficient or flexible. Notable contributions include the Galton tree-structure DEM [28], the Adams butterfly DEM [29] and the Schreier vector feedback DEM [30], used for first-order or restricted second-order shaping. The SMNR estimate is equally valid for all if H_{DEM} is known.

4. CLOCK JITTER AND SJNR ESTIMATE

A critical error source in high resolution DAC design is *clock jitter* or *timing* errors. Jitter errors can stem from noise and noise-coupling in on-chip and on-board clock circuitry [30]-[31], and in audio also from the digital

audio interface [32]. It is usually divided into two classes; wide-band random jitter caused by circuit noise and noise coupling in the clock generation circuitry and in-band sinusoid jitter caused by supply ripple and parasitic coupling from signal lines. Wide-band jitter can be white or pink PSD noise while in-band jitter can consist of uncorrelated or correlated sideband distortion. The presented estimates are limited to white PSD wide-band jitter and uncorrelated sideband jitter

4.1. Jitter Error Modelling

The jitter error waveform, being the real DAC output waveform minus the ideal (non-jittered) ditto, is illustrated for an otherwise ideal continuous-time DAC with hold reconstruction in fig.10. It will consist of a sum of error pulses and can for an N sample sequence be mathematically described as:

$$e_{jit}(t) = \sum_{n=0}^{N-1}\Big[y_a\big(nT\big) - y_a\big((n-1)T\big) \cdot \frac{1}{T}\Pi_{j(nT)}\left(nT + \frac{j(nT)}{2}\right) \cdot \operatorname{sign}\big(j(nT)\big)\Big] \qquad (23)$$

$\Pi_a(b)$ defines a rectangular window of unity height and width a, centred at b. The sampling period $T=1/f_s$. Considering *a single* error pulse on this form, for simplicity denoting its amplitude A and width J, taking the Fourier transform gives the spectrum:

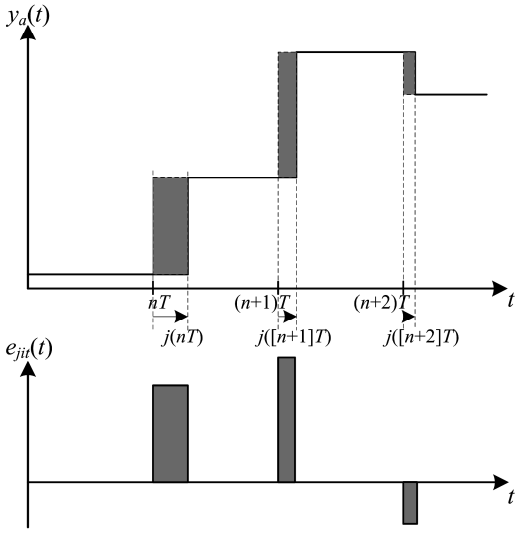


Figure 10: Jitter error waveform

$$\begin{aligned} \mathrm{E}_J(f) &= \int_{t=0}^{J} A\cdot e^{-\mathbf{i}\cdot 2\pi f\cdot t}\Big|_{J\geq 0} = -\int_{t=J}^{0} A\cdot e^{-\mathbf{i}\cdot 2\pi f\cdot t}\Big|_{J<0} \\ &= \frac{A\cdot\sin\big(2\pi.Jf\big)}{2\pi f}\cdot e^{-\mathbf{i}\cdot\pi.Jf} \\ &= A\cdot J\cdot\operatorname{sinc}\big(Jf\big)\cdot e^{-\mathbf{i}\cdot\pi Jf} \end{aligned} \qquad (24)$$

The sinc-function is by conventional definition given as:

$$\operatorname{sinc}\big(x\big) \overset{def}{=} \frac{\sin\big(\pi x\big)}{\pi x} \qquad (25)$$

The pulse spectrum expression (24) can be used to approximate the discrete Fourier transform (DFT) for a jittered sample sequence. The elaborate method [33] is to bandlimit it at $f_s/2$ with a brick-wall filter before sampling it into an error spectrum sequence $E_J(\omega,n)$, where $A=(y_a[n]-y_a[n-1])/T$ and $J=j(nT)=j[n]$ for all n. This can then be summed into a composite DFT for an N-sample sequence. The main drawback is that it requires massive computation time since a bandlimited sampled spectrum must be calculated for every n.

To get around the computation time problem, note that if $J\cdot f<<1$ the expression (24) approximates a constant:

$$E_J(f) \approx A\cdot J\big|_{Jf<<1} \qquad (26)$$

Assuming jitter in the picoseconds range, this is a good approximation up to many MHz. In oversampled DACs it's a reasonable and common approach [34]. The error waveform is now approximated as Dirac-pulses with weight given by the real pulse area $A\cdot J$, called *error area modelling*. Since there is now a constant $E_J(\omega,n)$ for each sampling instant, it becomes easy to compose DFT for the N-sample sequence:

$$\begin{aligned} E_{jit}\big(\omega\big) &= \sum_{n=0}^{N-1} E_J\big(\omega,n\big)\cdot e^{-\mathbf{i}\omega n} \\ &\approx \frac{1}{T}\sum_{n=0}^{N-1}\Big[\big(y_a\big[n\big]-y_a\big[n-1\big]\big)\cdot j\big[n\big]\Big]\cdot e^{-\mathbf{i}\omega n} \end{aligned} \qquad (27)$$

This approximation has a slight absence of high frequency drop-off and overshoot from the Gibbs effect on the error pulse, but since the pulse in real-life is strongly apodized by low-pass circuit effects the error area model is arguably as representative.

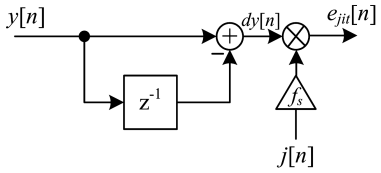


Figure 11: Error area model for jitter error simulation

Error area modelling is used for simulations in this paper as shown in fig.11.

4.2. SJNR estimation

From the model and (27), it is clear that there is time domain multiplication between the jitter sequence *j* and the differentiated DAC input sequence *dy*, implying frequency domain convolution:

$$S_{e_{jit}}(\omega) = \frac{1}{T^2}\left[S_{dy}(\omega) * S_{jit}(\omega)\right] \qquad (28)$$

This means that the jitter modulates the signal, causing sidebands if it is sinusoid and noise elevation if it is noise-like. Simulations in fig.12 of a jittered DSM confirm this. The underlying grey trace is without jitter. An audio range fifth-order DSM with L=64 is used.

Disregarding mismatch the DAC output *y* consists of a DC-offset $M/2$, a signal component *x* and the quantization noise. Differentiation removes the DC, while if *x* is a sinusoid at $\omega_x << \pi$ – a reasonable assumption for oversampled converters – then *dx* is also sinusoid at ω_x, with approximate amplitude $A_x \cdot \omega_x$. Both with a many-bit LPCM signal and a few-bit DSM REQ signal, in-band quantization noise is very low, meaning in-band jitter modulates only *dx*. If the jitter is also sinusoid, general expressions on the form $A \cdot \cos(\omega n)$ can be inserted for *dx* and *j* and via the trigonometric angle sum and difference identities the error is found to be sinusoids at $\omega_x \pm \omega_{jit}$ with amplitude:

$$A_{e_{jit}}^{(\omega_x \pm \omega_{jit})} = \frac{A_{jit} \cdot A_{dx}}{2T} \approx \frac{A_{jit} \cdot A_x \cdot \omega_x}{2T} \qquad (29)$$

Since A_{jit} and ω_x are of reciprocal magnitudes they can be normalized to T and f_s as normal, or to seconds and Hz. It is particularly noteworthy that in-band jitter distortion is identical in the LPCM-case and the DSM-case, and that the number of bits in the DSM won't make a difference. Intuitively jitter performance is worse with few bits because the output steps are much larger, but with in-band jitter this is not the case.

With wide-band white jitter, the convolution causes noise with a white PSD that depends on the entire spectrum of *dy*. Under the additive noise approximation, assuming unity STF, we have that:

$$S_{dy}(e^{i\omega}) \approx S_{dx}(e^{i\omega}) + \frac{1}{12 \cdot 2\pi}\left|\left(1 - e^{-i\omega}\right) \cdot H_{NTF}(e^{i\omega})\right|^2 \qquad (30)$$

For simplicity denoting the differentiated shaping function H_{dNTF}, total differentiated output power is:

$$\sigma_{dy}^2 \approx \sigma_{dx}^2 + \frac{1}{12}\left\|H_{dNTF}(e^{i\omega})\right\|_2^2 \qquad (31)$$

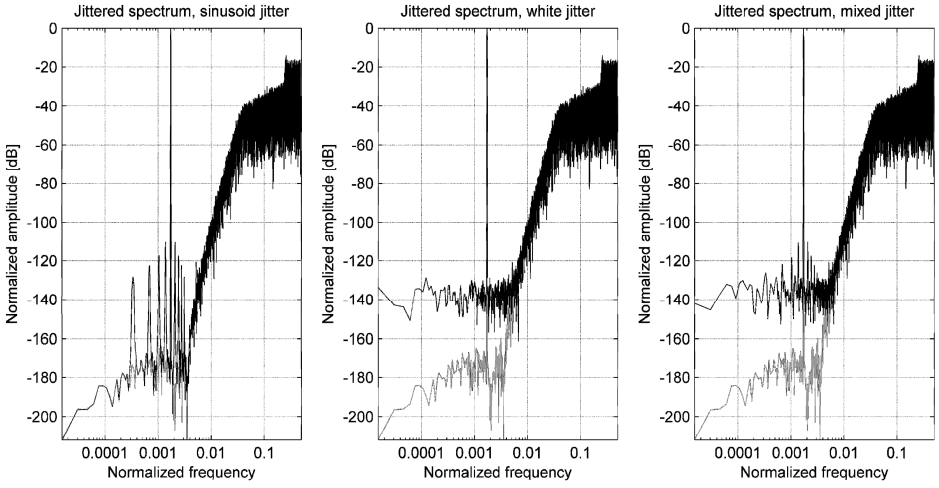


Figure 12: Simulated output spectra, DSM DAC with error area jitter model

If $S_{jit}(\omega)$ from (28) is white, the error PSD is also white. The total error power is given by the multiplication of the two and in-band error power is thus:

$$\sigma_{e_{jit}}^2 \approx \frac{\sigma_{dy}^2 \cdot \sigma_j^2}{T^2 \cdot L} \tag{32}$$

For sinusoid input the SJNR can thus be estimated as:

$$SJNR \approx 10 \cdot \log_{10}\left(\frac{(k \cdot M)^2 \cdot T^2 L}{\left((k \cdot M)^2 \cdot \omega_x^2 + \frac{2}{3} \cdot \left\| H_{dNTF}\left(e^{i\omega}\right) \right\|_2^2 \right) \cdot \sigma_j^2} \right) \tag{33}$$

Note that L doubles if T is halved, meaning that with fixed jitter variance, the SJNR decreases 3dB per octave OSR. But for typical oscillators the phase noise standard deviation is inversely proportional to T [35] and then the SJNR is not affected by OSR.

As long as quantization noise dominates (33), the denominator in the expression is relatively constant and the SJNR increases 6dB per bit. With very little quantization noise power, e.g. in high-res LPCM, white jitter SJNR will converge towards T^2L/σ_j^2.

4.3. SJNR estimate verification

The SJNR estimates were verified with functional simulations using the error area model. The modulator is a fifth-order DSM for audio with 44.1kHz f_{s_in} and L=64. Figure 13 shows the estimate and simulation results with sinusoid in-band jitter.

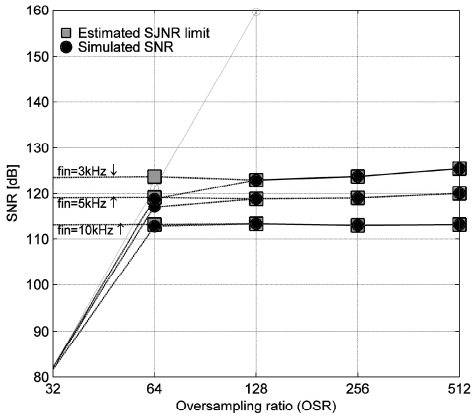


Figure 13: Estimated SJNR limit and simulated SNR; 1-bit DSM, 50ps 1kHz jitter

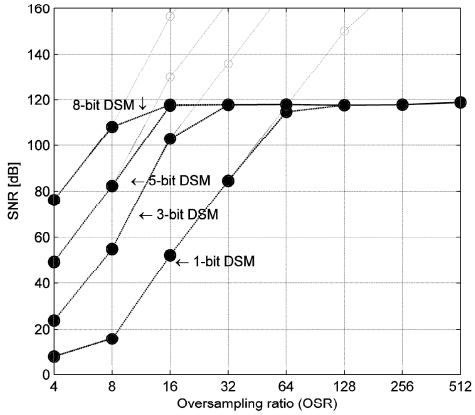


Figure 14: Simulated SNR of different DSM DACs; 50ps, 1kHz jitter, f_{in}=5kHz

Figure 14 shows SNR simulations with sinusoid in-band jitter, using different numbers of levels in the REQ. It is confirmed that the SJNR limitation is for this type of jitter is independent of the OSR as well as the number of levels in the DSM REQ.

Figure 15 compares estimates and simulation results for white PSD jitter noise. Simulations are swept over different OSR, different number of bits and different NTF aggressiveness. The estimation method is confirmed to be accurate and it is seen that a few-level DSM REQ will severely compromise the SJNR limit. For many-bit DSMs, significant SJNR gains also can be made from designing the DSM conservatively. This means using an overly aggressive NTF to maximize SQNR will compromise SJNR. Thus the DSM should only be aggressive enough for sufficient SQNR and a trade off established for the expected wide-band jitter noise from the clock generation. The SQNR is shown in the greyed out traces.

As a final note it should be emphasised that while white jitter is a product of the converter's clock circuitry and will reduce its SNR, in-band tonal jitter often stems external sources including the (audio) interface. Since it will not appear in measurements of a DAC's intrinsic SNR, it is often regarded as a critical "hidden" error source. Although we have established that the DSM topology will not affect a converter's susceptibility to such jitter, converters may have different *jitter transfer functions* (JTF) from the interface to the sampling or reconstruction point. The JTF will depend on the design and the implementation quality of the clock recovery circuit. Test methods to specifically address the in-band tonal jitter susceptibility of a DAC's clock recovery circuitry do exist, notably the "J-test" [36].

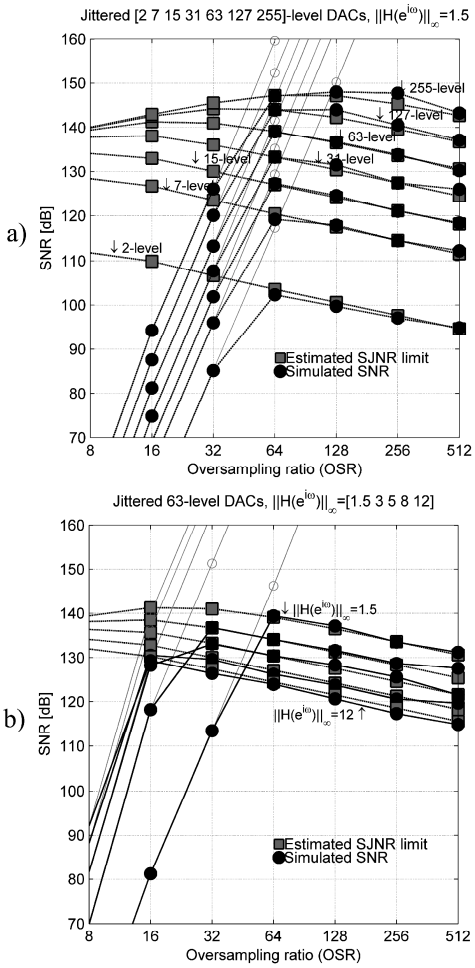


Figure 15: SJNR estimate vs. simulated SNR, 5ps RMS white random jitter, a) different number of levels, b) 63-level and different NTF aggressiveness

5. DYNAMIC ERRORS AND INTER-SYMBOL-INTERFERENCE

Another topology dependent error source in a DAC is signal dependent errors from non-ideal DAC element switching [37]-[38], commonly called inter-symbol interference (ISI). Switching errors or glitches can stem from charge injection and clock feedthrough causing overshoot, or time constants in the switching network causing limited rise and fall times. DAC ISI should not be confused with ISI data errors in digital communications, although the physical nature is similar.

5.1. ISI error modelling and estimation

The generalized output from a single DAC element is shown in fig.16. Using the same error area approach as for the jitter error, one can assume an on-error $-e_{on}$ or an off-error e_{off} added to each element when switched on or off. This means that as for jitter, one can find a discrete error area sequence.

Element i is switched between 0 and w_i, but for simplicity it's assumed mismatch-free and switched between 0 and 1. In a thermometer encoded DAC without DEM, K elements are switched on if the DAC input value increases with K from the previous to the current sample. If it decreases by K, then K elements are switched off. This leads to the general ISI error expression:

$$e_{ISI_{term}}[n] = \begin{cases} \left(y[n]-y[n-1]\right)\cdot e_{off}, & \left(y[n]-y[n-1]\right)<0 \\ \left(y[n]-y[n-1]\right)\cdot e_{on}, & \left(y[n]-y[n-1]\right)\geq 0 \end{cases} \quad (34)$$

With DWA the situation is quite different. If the last two samples combined are less or equal to the total number of elements M, then $y[n]$ elements are turned on and $y[n\text{-}1]$ elements are turned off. If they are larger the modulo pointer wraps, meaning $M\text{-}y[n\text{-}1]$ elements are turned on and $M\text{-}y[n]$ elements are turned off.

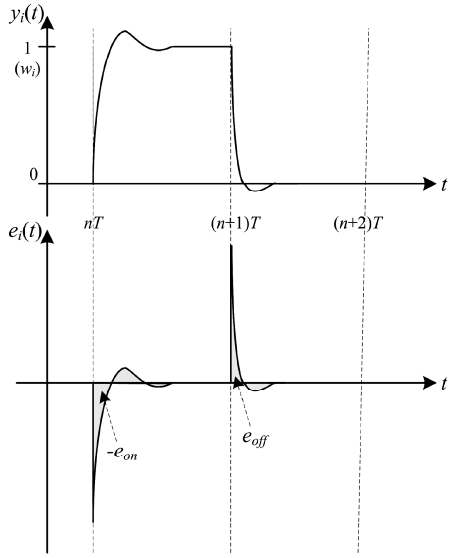


Figure 16: ISI error contribution from a single switched DAC element w_i

This leads to the general ISI error expression:

$$e_{ISI_{DWA}}[n] = \begin{cases} y[n]\cdot e_{on} - y[n-1]\cdot e_{off} & , y[n]+y[n-1] \le M \\ (M - y[n-1])\cdot e_{on} - (M - y[n])\cdot e_{off} & , y[n]+y[n-1] > M \end{cases} \quad (35)$$

Without DEM it is clearly seen that if e_{on}=e_{off} the error waveform is a linear function of dy and hence harmless. If switching is asymmetric the error waveform is asymmetric and hence consists of even harmonics. Since y is generated by a DSM REQ it's not possible to predict analytically.

Taking advantage of the additive quantization noise estimate to use superposition, if we first evaluate dx and assume x is a sinusoid of amplitude A_x and frequency ω_x, Fourier series expansion yields [37]:

$$A_{ISI_{term}}^{(k\cdot\omega_x)} \approx \frac{2\cdot\left|e_{off} - e_{on}\right|}{\pi(k+1)(k-1)}\cdot A_x\cdot\omega_x \quad k = 2,4,6... \quad (36)$$

Evaluating only the shaped noise term, simulations have shown the asymmetry whitens the noise and therefore the signal-to-switching-noise-ratio (SSNR) will be:

$$SSNR = 10\cdot\log_{10}\left(\frac{(k\cdot M)^2\cdot L}{\frac{2}{3}\cdot\left\|NTF_d\right\|_2^2\cdot\left(e_{off} - e_{on}\right)^2}\right) \quad (37)$$

As such we have separate estimates for the reduction in both SNR and spurious free dynamic range (SFDR), caused by asymmetric switching.

It is seen from (35) that the error term with DWA depends directly on y rather than dy. Again disregarding quantization noise; if y=M/2+x and x is sinusoid, equal e_{on} and e_{off} again leads to e_{ISI} being a linear product of x, seeing how each half-period satisfies either the first or second condition of (35). If e_{on} and e_{off} are different e_{ISI} will again be an asymmetric sinusoid and contain harmonics. Fourier expansion of e_{ISI} with DWA [37] will produce the same result as (25), but with even harmonics proportional to A_x rather than A_{dx}:

$$A_{ISI_{DWA}}^{(k\cdot\omega_x)} = \frac{2\cdot\left|e_{off} - e_{on}\right|}{\pi(k+1)(k-1)}\cdot A_x \quad k = 2,4,6... \quad (38)$$

Additional in-band noise caused by the shaped noise component ε_{dsm}, is from (35) seen to be non-linearly filtered and is not as easily predictable. Since strong harmonics is by far the most severe switching error effect in a DWA DAC, estimates for in-band noise elevation have not been pursued. Simulations show however that in-band noise does increase, but moderately and with a white noise floor.

Making ISI estimates for other DEM schemes will be an exercise that depends entirely on how the DEM algorithm switches elements. For DEM with higher order shaping it would be very difficult to predict sample-to-sample element switching, but since all high-pass shaping DEM schemes are based on equal element usage, the ISI error is expected to be dominated by even harmonics like DWA.

5.2. ISI simulations and estimate verification

To determine the on and off switching errors areas of a DAC element accurately, will require a transistor level simulation of its circuitry. It can however in many cases be assumed to be dominated by slewing, causing the element to have limited rise and fall times. To make the area calculation simple we assumed linear slewing as in [37], with limited rise- and fall-times t_r and t_f so that the switching errors became:

$$e_{on} = \frac{1}{2}\cdot\frac{t_r}{T},\ e_{off} = \frac{1}{2}\cdot\frac{t_f}{T} \quad (39)$$

Figure 17 shows the simulated output spectra of DEM switched and otherwise ideal DACs having 5ps slewing asymmetry. The DAC input is generated by a fifth-order 5-bit DSM REQ for audio, L=64 and f_{s_in}=44.1kHz. The "x"-marks show estimated harmonic components according to (36) and (38). It is seen that without DEM the harmonics are buried in the noise floor, while with DWA harmonics are severe. The estimate is accurate up to at least HD6 and very accurate for HD2 and HD4. It is not known why odd harmonics appear, but HD2 is clearly limiting the SFDR. With R2DWA, the spectrum is more benign with a higher noise-floor and less harmonics. Still HD2 is significant.

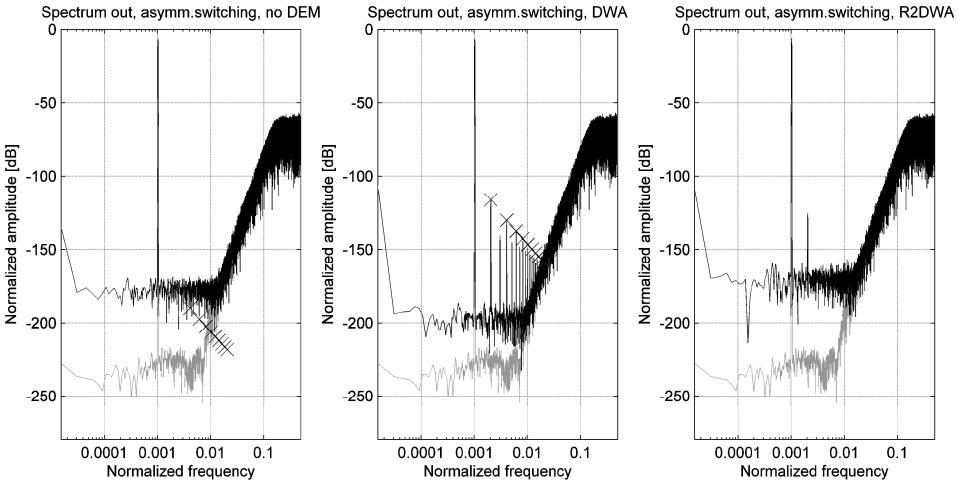


Figure 17: Simulated output spectra; DSM DAC with error area ISI model, 5ps asymmetry

Figure 18 compares the estimates for SSNR without DEM (37) and harmonic distortion with DWA to simulated results. It is seen that the estimates can effectively predict SSNR and SFDR from asymmetric switching. With constant t_r and t_f it's understood from (39) why the ISI-limited SNR decreases by 3dB per octave OSR and SFDR by 6dB.

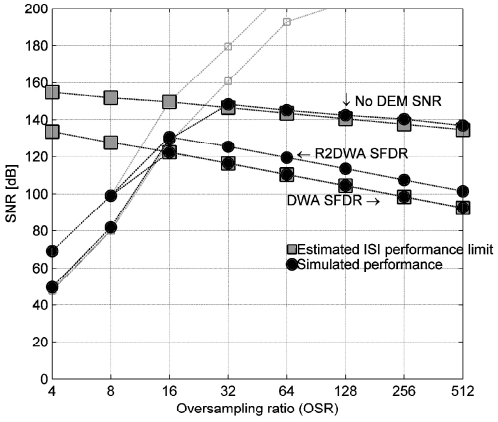


Figure 18: Estimated ISI performance vs. simulated SNR; DSM DAC with 5ps switching asymmetry

To minimize ISI one should preferably avoid DEM, but this is not possible in a multi-bit hi-res DAC. It is seen from (37) that a large M is desirable. Reduced activity DWA to combat ISI distortion has also been proposed [39]. ISI can also be eliminated with Return-to-Zero coding [40], which resets every element within a sample period so that each element's switching error contribution becomes the linear function:

$$e_{ISI_{RZ}}[n] = y[n] \cdot \left(e_{off} - e_{on}\right) \qquad (40)$$

Unfortunately RZ also reduces the signal energy, increases the switching speeds and makes the DAC highly jitter susceptible. The RZ waveform is reset to zero for every sample, meaning wideband jitter errors will be at the 1-bit level no matter how many bits are used in the DSM REQ. To alleviate this, waveform-preserving RZ techniques like dual-RZ [41] and time-interleaving [42] have been proposed, but at the cost of higher analog complexity. An ingenious solution for exploiting pulse width modulation (PWM) to eliminate both mismatch and ISI without output waveform distortion was recently published [43]. Its main disadvantage is that for a B-bit DAC, the PWM time resolution must be $2^B \cdot L$ times the input sampling rate

6. CONCLUSIONS

This paper has presented an overview of error sources of particular interest to high-res delta-sigma DA design, and estimates for how these depend on the modulator's NTF and number of bits. It is shown how the additive noise DSM approximation can be extended from only predicting quantization errors, to also predict mismatch errors, dynamic errors and jitter errors as part of the performance assessment. These errors are likely to be dominant in a real-world DAC implementation and it is therefore of high importance to error budgeting and modulator design tradeoffs that they are included.

Significant error sources not reviewed here include analog thermal and shot noise [44], as well as finite output impedance causing INL when the DAC is driving a load [45]. These errors are circuit dependent and not waveform dependent, therefore they won't relate to the DSM in the same way.

Neither the presented simulation models nor the estimates based on the additive noise model are intended to or should replace transistor level analog circuit simulations. The high-level Matlab models do ignore some lower order effects such as time skew between DAC elements [46] and graded mismatch [47], which may cause inaccuracies. The estimates do however provide quick and easy methods to find initial DSM parameters such as OSR, NTF and the number of bits in the REQ based on a target performance specification. The information provided should also enable the reader to quickly make qualified error predictions and DSM parameter trade-offs, based on known physical properties in the analog circuit design.

7. ACKNOWLEDGEMENTS

This work was supported by the Norwegian Research Council under grant 162101 SPECK.

8. REFERENCES

[1] G.E.Moore; "Cramming More Components Onto Integrated Circuits", *Electronics Magazine*, vol.38, no.8, 1965.

[2] H.Inose, Y.Yasuda and J.Marakami; "A Telemetering System by Code Modulation, Delta-Sigma Modulation", *IRE Trans. Space, Electronics and Telemetry*, SET-8, pp. 204-209, Sept.1962.

[3] H.Inose and Y.Yasuda; "A Unity Bit Coding Method by Negative Feedback", *IEEE Proceedings*, Vol. 51, pp. 1524-1535, Nov.1963.

[4] D.J.Goodman, "The Application of Delta Modulation of Analog-to-PCM Encoding", *Bell System Tech.J.*, vol.48, pp.321-343, Feb.1969.

[5] G.R.Ritchie, J.C.Candy, and W.H.Ninke; "Interpolative Digital-to-Analog Converters", *IEEE Trans. Communications*, Vol. COM-22, pp. 1797-1806, Nov.1974..

[6] Philips Intellectual Property and Standards, "IEC-908: Compact Disc Digital Audio – The Red Book", *International Electrotechnical Commission Standards Document*, no. 28/10/04-3122 783 0027 2, Jun.1980.

[7] R.W.Adams; "Design and Implementation of an Audio 18-Bit Analog-to-Digital Converter Using Oversampling Techniques", *J. Audio Eng. Soc.*, Vol. 34 No.3, pp. 153-166, March 1986.

[8] W.R.Bennett; "Spectra of Quantized Signals", *Bell Systems Tech. J.*, vol.27, 1948.

[9] B.Widrow; "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory", *IRE Trans. Circuit Theory*, vol. CT-3, pp. 266-276, Dec. 1956

[10] S.P.Lipshitz, R.A.Wannamaker and J.Vanderkooy; "Quantization and Dither: A Theoretical Survey", *J. Audio Eng. Soc.*, Vol. 40, No. 5, pp. 355-375, May 1992

[11] R.M.Gray; "Quantization Noise Spectra", *IEEE Trans. Information Theory*, vol. 36, no. 6, Nov. 1990

[12] R.Schreier; "An Empirical Study of High-Order Single Bit Delta-Sigma Modulators", *IEEE Trans. Circuits and Systems - Part II*, vol. 40, pp 461-466, Aug.1993

[13] J.Candy; "The Structure of Quantization Noise from Delta-Sigma Modulators", *IEEE Trans. Communications*, vol.29, no.9, 1981

[14] S.R.Norsworthy and D.A.Rich; "Idle Channel Tones and Dithering in Delta-Sigma Modulators", *Audio Eng. Soc. Convention Paper 3711*, 95th Convention New York, 7-10 October 1993

[15] I.Løkken, A.Vinje and T.Sæther; "Noise Power Modulation in Dithered and Undithered High-Order Sigma Delta Modulators", *J. Audio Eng. Soc.*, vol.54, pp.841-854, September 2006

[16] W.Chou et.al; "Multistage Sigma-Delta Modulation", *IEEE Trans. Information Theory*, vol.35, no 4, July 1989

[17] H.Kato; "Trellis Noise-Shaping Converters and 1-bit Digital Audio", *Audio Eng. Soc. Convention Paper 5615*, 112th Convention Munich, May 2002

[18] W.L.Lee; "A Novel Higher-Order Interpolative Modulator Topology for High Resolution Oversampling A/D converters", *M.Sc. thesis, MIT, Cambridge, MA*, June 1987.

[19] J.Reiss; "Towards a Procedure for Stability Analysis of High Order Sigma Delta Modulators", *Audio Eng. Soc. Convention Paper 6549,* 119th Convention New York, October 2005

[20] M.J.M. Pelgrom et al., "Matching Properties of MOS Transistors", *IEEE J. Solid-State Circuits*, vol.24, no.5, pp. 1433-1439, October 1989

[21] R.J.Van De Plassche; "Dynamic Element Matching for High Accuracy Monolithic DA Converters", *IEEE J. Solid State Circuits*, vol.11, no.6, pp.795-800, Dec.1976.

[22] L.R.Carley; "A Noise Shaping Coder Topology for 15+ bit Converters", *IEEE J. Solid State Circuits*, vol.24, no.2, pp.267-273, April 1989.

[23] R.T.Baird, T.S.Fiez; "Linearity Enhancement of Multi-Bit Δ-Σ A/D and D/A Converters Using Data Weighted Averaging", *IEEE Trans. Circuits & Systems II: Analog and Digital Signal Processing*, vol.42, no.12, pp.753-762. Dec. 1995.

[24] M.Vadipour; "Techniques for Preventing Tonal Behaviour of Data Weighted Averaging Algorithm in ΔΣ-Modulators", *IEEE Trans. Circuits and Systems II*, vol.47, no.11, pp 1137-1144, Nov.2000

[25] R.K.Henderson and O.Nys; "Dynamic Element Matching Techniques with Arbitrary Noise Shaping Function", *Proc. IEEE Int. Symp. Circuits and Systems ISCAS'96*, pp.293-296, May 1996

[26] Xue-Mei Gong; **"**An Efficient Second-Order Dynamic Element Matching Technique for a 120 dB Multi-Bit Delta-Sigma DAC", *Audio Eng. Soc. Convention Paper 5124*, 108th Convention of the AES, Paris, February 2000.

[27] I.Galton; "Spectral Shaping of Circuit Errors in Digital-to-Analog Converters", *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing,* vol. 44, no. 10, pp.808-817, Nov., 1997.

[28] R.Adams, K.Nguyen, K.Sweetland; "A 113dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling", *IEEE ISSCC Dig. of Tech Papers*, vol.41, pp.62-63, Feb.1998.

[29] R.Schreier, B.Zhang; "Noise-Shaped Multibit D/A Converter Employing Unit Elements", *Electronic Letters*, vol.31, no.20, pp 1712-1713, Sept. 1995.

[30] P.Heydari; **"**Analysis of the PLL Jitter Due to Power/Ground and Substrate Noise", *IEEE Trans. Circuits and Systems I: Regular Papers*, vol.51, no.12, pp.2404–2416, Dec.2004.

[31] J.L.LaMay, H.T.Bogard; "How to Obtain Maximum Practical Performance from State of the Art Delta-Sigma Analog to Digital Converters", *9th IEEE Instrumentation and Measurement Technology Conference*, 1992, IMTC '92, pp.552-560, May 1992

[32] C.Dunn, M.O.J.Hawksford; "Is the AESEBU/ SPDIF Digital Audio Interface Flawed?", *Audio Engineering Society Convention Paper 3360,* AES 93rd Convention, San Francisco, October 1992.

[33] M.O.J.Hawksford: "Jitter Simulation in High Resolution Digital Audio", *Audio Eng. Soc. Convention Paper 6864*, 121st Convention San Francisco, October 2006.

[34] K.Doris, A.van Roermund, D. Leenaerts; "A General Analysis on the Timing Jitter in D/A Converters", *Proc. IEEE Int. Symp. Circuits and Systems, ISCAS 2002*, vol.1, pp.117-120, May 2002

[35] J.A.McNeill; "Jitter in ring oscillators", *IEEE J. Solid State Circuits*, vol.32, no.6, pp.870-879, June 1997.

[36] J.Dunn; "Jitter: Specification and Assessment in Digital Audio Equipment", *Audio Eng. Soc. Convention Paper 3361*, AES 93rd Convention, San Francisco, October 1992.

[37] M.Clara, A.Wiesbauer,W.Klatzer; "Nonlinear Distortion in Current-Steering D/A-Converters Due to Asymmetrical Switching", *Proc. IEEE Int. Symp. Circuits and Systems*, vol.1, pp.285-288, May 2004.

[38] K.O.Andersson, J.J.Wikner; "Characterization of a CMOS Current-Steering DAC Using State-Space Models", *Proc. IEEE 2000 Midwest Symp. Circuits and Systems*, vol. 2, pp.668-671, Aug.2000

[39] A.Bicakci, G.Singh: "A ΔΣ DAC with Reduced Activity Data Weighted Averaging and Anti-Jitter Digital Filter", *Proc. IEEE 2005 Custom Integrated Circuits Conference*, pp.383-386, Sept.2005.

[40] B.P.Del Signore et.al.; "A Monolithic 20-b Delta-Sigma A/D Converter", *IEEE J. Solid-State Circuits*, vol.25, no.6, pp.1311-1317, Dec.1990

[41] R.Adams, K.Ngyuen, K.Sweetland; "A 112 dB SNR Oversampling DAC with Segmented Noise-shaped Scrambling", *AES Convention Paper 4774*, 105th Convention Audio Eng. Soc., September 1998

[42] M.Clara, W.Klatzer, A.Wiesbauer, D.Straeussnigg; "A 350MHz Low-OSR Delta-Sigma Current-Steering DAC With Active Termination in 0.13µm CMOS", *IEEE Int. Solid-State Circuits Conference*, ISSCC Digest of Technical Papers, pp.118-588, February 2005

[43] D.Reefman et.al; "A New Digital-to-Analogue Converter Design Technique for HiFi Applications", *AES Convention Paper 5846*, 114th Convention Audio Eng. Soc., March 2003

[44] R.Schreier, G.Themes; "Understanding Delta-Sigma Data Converters", *John Wiley & Sons*, ISBN 0-471-46585-2, ch.9.6.3 "DAC and Reconstruction Filter Design", pp.355-357, 2005

[45] D.Mercer; "A Study of Error Sources in Current Steering Digital-to-Analog Converters", *IEEE Custom Integrated Circuits Conference*, pp.185-190, October 2004

[46] K.Doris, D.M.W.Leenaerts, A.H.M.van Roermund; "Time Non Linearities in D/A Converters", *European Conference on Circuit Theory and Design*, Helsinki Finland, pp.III-353-III-356, Helsinki, August 2001.

[47] K.O.Andersson, J.J.Wikner; "Modeling of the Influence of Graded Element Matching Errors in CMOS Current-Steering DACs", *Proc. 17th Norchip Conference*, Oslo Norway, November 1999.

Appendix 7

Paper V:

I. Løkken, A. Vinje, T. Sæther, "Delta-Sigma DAC Topologies for Improved Jitter Performance", *Audio Eng. Soc. Convention Paper 7497,* 124th Convention of the Audio Eng. Soc. – Discover New Horizons in Audio, Amsterdam NL, (2008 May).

© 2008 AES. Reprinted with permission.

Audio Engineering Society

Convention Paper 7497

Presented at the 124th Convention
2008 May 17–20 Amsterdam, the Netherlands

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Delta-Sigma DAC Topologies for Improved Jitter Performance

Ivar Løkken, Anders Vinje, and Trond Sæther

Norwegian University of Science and Technology, Trondheim, Norway
ivar.loekken@iet.ntnu.no
anders.vinje@iet.ntnu.no
trond.saether@iet.ntnu.no

ABSTRACT

Specifications for audio digital-to-analog converters (DACs) place requirements on the analog circuit design that contradict physical design conditions in a modern, digital-oriented system on a chip process. Because of low supply voltages, use of current-steering DACs has become the dominant choice for high resolution applications. Fed by a delta-sigma modulator that requantizes the digital signal to a manageable number of bits, the current-steering DAC is a continuous time type converter without any discrete time filtering. This makes it very susceptible to sampling clock jitter. In this paper, jitter induced distortion is addressed at a topology level, investigating design choices for the delta-sigma requantizer and the possible use of semidigital multi-bit current-steering filter DACs to reduce problems with jitter susceptibility.

1. INTRODUCTION

Physical restraints in design of analog and mixed-mode integrated circuits (ICs) make it challenging to obtain the same rate of performance improvement for audio digital-to-analog converters (DACs) as for the digital audio processors feeding them. With speeds increasing and feature sizes decreasing in more and more digital oriented IC technologies, consequent requirements for lowered supply voltages and power losses make it more difficult to integrate analog circuitry with high audio-grade resolution. In particular, sampling clock jitter is becoming a dominant source of distortion. Previously it was normal to use voltage-mode DACs with switched capacitor filtering to reduce the effects of clock jitter. In modern sub-micron technologies with supply voltages in the 1V range, it is almost impossible to implement op-amps - and thus switched capacitor filters - with sufficiently low noise and distortion for high resolution audio. This has led to most modern DACs being realized as continuous time current-steering converters, with external I-V conversion using a dedicated high-voltage op-amp chip or a discrete transistor stage.

Since current-steering converters have no low-pass filtering before the discrete-to-continuous time interface, clock jitter becomes a particularly perilous performance bottleneck. The high amount of out-of-band noise generated by the delta-sigma modulating

requantizer (DSM REQ) will also further compromise the jitter performance. This has led to the necessity of using multi-bit DSM REQs with DAC mismatch errors shaped through dynamic element matching (DEM). The fast increasing complexity of DEM schemes does however limit the number of bits in a multi-bit DSM REQ to only a few. One way to address this is by segmenting the DAC, another is to use a semidigital finite impulse response (FIR) DAC that filters the DSM output. Both will be addressed in this paper.

Most analysis behind the paper is done in the sampled domain, even though jitter is a continuous time problem. The first section reviews how jitter distortion can be approximated as a sample-by-sample error and hence simulated in a fast discrete time simulator. The discrete time signal's angular frequency ω is defined as:

$$\omega \overset{def}{=} \frac{2\pi f}{f_s} \tag{1}$$

In this definition f_s is the DAC sampling frequency *including oversampling*, or $f_s=f_{si}\cdot OSR$ where OSR is oversampling ratio and f_{si} is the system input sampling rate (typ. 44.1kHz for CD). f is the frequency in Hz. This notation is used throughout.

2. JITTER DISTORTION MODELLING IN THE AUDIO DAC

The jitter error waveform, defined as the real output waveform minus the non-jittered output waveform, is illustrated for an otherwise ideal continuous time DAC with hold reconstruction in fig.1. It can be mathematically described from:

$$\begin{aligned}\varepsilon_{J_n}(t) = &\left(y(nT) - y((n-1)T)\right) \\ &\times \frac{1}{T}\Pi_{J(nT)}\left(nT + \frac{j(nT)}{2}\right) \\ &\times \operatorname{sign}\left(j(nT)\right)\end{aligned} \tag{2}$$

In (2), $\Pi_a(b)$ defines a rectangular window of unity height and width a that is centred at b. The sampling period is given as $T=1/f_s$. Considering *a single* error pulse on the form (2), for simplicity denoting its amplitude A and its width J; computing the Fourier transform will give a pulse spectrum as described in the following equation:

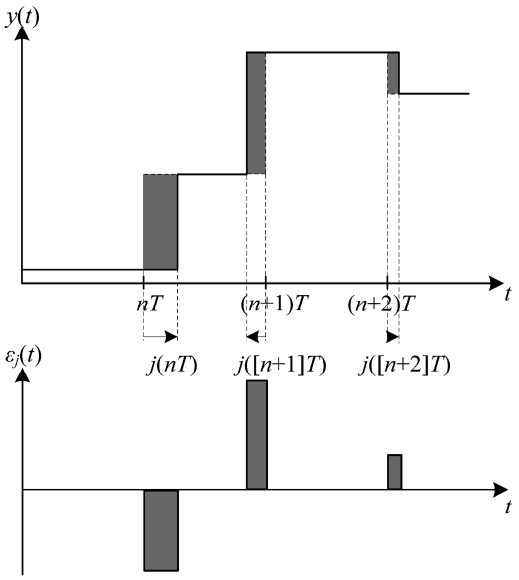


Figure 1 - Jitter error waveform

$$\begin{aligned}\hat{\mathrm{E}}_J(f) &= \int_{t=0}^{J} A\cdot e^{-\mathbf{i}\cdot 2\pi f\cdot t}\Big|_{J\geq 0} = -\int_{t=J}^{0} A\cdot e^{-\mathbf{i}\cdot 2\pi f\cdot t}\Big|_{J<0} \\ &= \frac{A\cdot\sin(2\pi Jf)}{2\pi f}\cdot e^{-\mathbf{i}\cdot\pi Jf} \\ &= A\cdot J\cdot\operatorname{sinc}(Jf)\cdot e^{-\mathbf{i}\pi Jf}\end{aligned} \tag{3}$$

The sinc-function is by conventional definition given by the expression:

$$\operatorname{sinc}(x) \overset{def}{=} \frac{\sin(\pi x)}{\pi x} \tag{4}$$

The pulse spectrum expression (3) can be used to approximate the Discrete Fourier Transform (DFT) for a jittered sample sequence. In Hawksford's paper [1], it is bandlimited at $f_s/2$ with a brick-wall filter before being sampled into an error spectrum set $\hat{E}_J(\omega,n)$, where $A=(y[n]-y[n\text{-}1])/T$ and $J=j(nT)$ for all n. This set can then be summed into a composite DFT from n=0 to N-1 for an N-sample sequence. The main drawback of this

approach is that it requires massive computation time since a bandlimited, sampled spectrum must be calculated for each n before being combined into the total DFT. The brick-wall filter can also not be ideal and in a real, physical world it is likely that the jitter pulses are deformed by physical band limits in the system.

To get around the computation time problem, note that if $J \cdot f << 1$ the expression (3) approximates a constant:

$$\hat{E}_J(f) \approx A \cdot J\big|_{Jf<<1} \tag{5}$$

Assuming jitter in the picoseconds range, (5) is a good approximation up to many MHz. In oversampled audio DACs it's a reasonable and common approach. Physically the error waveform is now approximated as Dirac-pulses with weight given by the real pulse area $A \cdot J$, called *error area modelling*. Since there is now a constant $\hat{E}_J(\omega,n)$ for each sampling instant n, it becomes easy to sum the DFT for an N-sample sequence:

$$\begin{aligned} E_J\left(e^{i\omega}\right) &= \sum_{n=0}^{N-1} \hat{E}_J(\omega,n) \cdot e^{-i\omega n} \\ &\approx \frac{1}{T} \sum_{n=0}^{N-1} \left[\left(y[n] - y[n-1] \right) \cdot j[n] \right] \cdot e^{-i\omega n} \end{aligned} \tag{6}$$

The difference between using (3) and (5) is a slight absence of high frequency drop-off and some overshoot due to the Gibbs effect from the error pulse. Since it in real-life is strongly apodized by low-pass circuit effects, the error area model is arguably as representative and requires *much* less computation time. Error area modelling is used for simulations in this paper.

Clock jitter causing the errors can stem from the clock regeneration circuitry in the DAC [2] or from the digital audio interface [3]. It can be both uncorrelated and correlated with the audio content [4]. It is generally divided in two classes; baseband and wideband jitter [5]. We define the DAC input signal y as $y=x+q$, where x is the (audio) signal term and q is the digital (quantization) noise. The error appears as modulation between the DAC input spectrum and the jitter spectrum as shown in fig.2. In a high-res LPCM system with little quantization noise $\sigma_q^2 << \sigma_x^2$ and $y \approx x$, therefore the jitter error appears as mixing with x only making it equivalent to regular sampling jitter [4].

A DSM REQ on the other hand may have significant quantization noise, especially if it has a few bit output. Its time domain output sequence $y[n]$ is difficult or impossible to predict analytically [6], but a very much used approximation in the frequency domain is that its power spectral density (PSD) is given by:

$$\begin{aligned} \left|Y\left(e^{i\omega}\right)\right|^2 &= \left|X\left(e^{i\omega}\right)\right|^2 + \left|Q\left(e^{i\omega}\right)\right|^2 \\ &= \left|X\left(e^{i\omega}\right)\right|^2 + \frac{1}{12 \cdot 2\pi} \cdot \left|H_{NTF}\left(e^{i\omega}\right)\right|^2 \end{aligned} \tag{7}$$

It assumes a unity in-band signal transfer function (STF) and q being an independent noise source with constant power 1/12 normalized to the quantizer step size (LSB). $H_{NTF}(e^{i\omega})$ is the noise transfer function (NTF) [6]-[7].

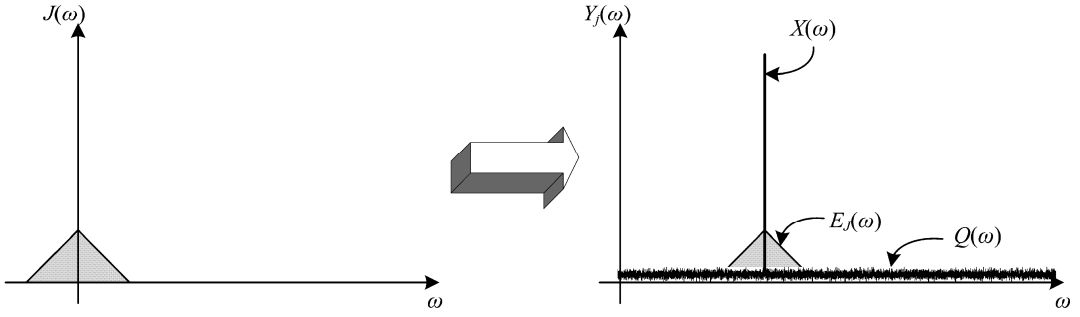


Figure 2 - Resulting distortion from a jitter source in the frequency domain

The NTF is chosen so that the shaped quantization noise has very little in-band energy, meaning that in-band it approximates a high resolution LPCM signal.

From (6) it is given that the jitter error spectrum as mentioned appears as a convolution, or more informally modulation, between the jitter spectrum and the spectrum of the differentiated DAC input. This means that baseband jitter creates additional in-band error content by modulating the signal and in-band quantization noise only. Since in-band quantization noise is very small, distortion from baseband jitter is thus equivalent to the high resolution LPCM case. Wideband jitter on the other hand, creates additional in-band error energy from modulation with the *entire* differentiated DSM spectrum, including out-of-band quantization noise. It therefore largely depends on the DSM REQ topology; immunity to wideband jitter can be improved substantially by increasing the number of bits or using a conservative NTF to lower out-of-band quantization noise. It can also be reduced with switch-cap filtering which suppresses out-of-band noise while in discrete time, but because of the need for huge on-chip capacitors switch-cap for high resolution is receding in modern sub-micron CMOS.

Wideband jitter is largely caused by noise in the digital processing and clock generation circuitry. Designing clock circuitry with low enough jitter for very high resolution audio is difficult and power-consuming. Therefore the use of multi-level DSM REQs has become the de facto design approach in high-res audio DACs [8]. The main drawback is that a multi-level DAC requires mismatch-shaping DEM algorithms [9] to avoid severe performance degradation from element mismatch [10]. For all known DEM algorithms the complexity increases very rapidly with the number of levels [11], limiting most published state-of-the-art DACs to five bits or less REQ. Because of jitter susceptibility future state-of-the-art DACs are likely to need more bits than the current art.

3. WIDEBAND JITTER SUSCEPTIBILITY AND THE NEED FOR MORE BITS

If the jitter j is a white random process independent of y, the convolved jitter error will also be a white random process. Its total power will be given by the product of the convolution terms' power. The in-band part of this is $1/OSR$ of the total, meaning in-band error power from uncorrelated white jitter is:

$$\sigma_{\varepsilon_j}^2 = \frac{\sigma_j^2 \cdot \sigma_{dy}^2}{OSR \cdot T^2} = \sigma_j^2 \cdot \sigma_{dy}^2 \cdot OSR \cdot f_{si}^2 \tag{8}$$

The notation dy is used for differentiated y, meaning $dy[n]=y[n]-y[n\text{-}1]$ for all n. If the DAC input signal y is made up of a sinusoid x that is quantized by a DSM REQ - its peak-to-peak amplitude A and frequency ω_x - the total power of dy from (7) is approximately:

$$\sigma_{dy}^2 \approx \frac{A^2}{8}\omega_x^2 + \frac{1}{12}\left\|H_{dNTF}\right\|_2^2 \qquad , \left\|H\right\|_2 \overset{def}{=} \sqrt{\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left|H\left(e^{i\omega}\right)\right|^2 d\omega} \tag{9}$$

The differentiated x is also a sinusoid at ω_x, its approximate amplitude $A\cdot\omega_x$ as long as $\omega_x<<1$. The total power of q is found by integrating its PSD over the whole frequency range π to π. The notation $dNTF$ is used for the differentiated NTF.

For M-level DSM REQs, the maximum peak-to-peak amplitude A is roughly proportional to M, or $A_{max}=k\cdot M$ where k is a constant. The maximum signal-to-jitter-noise ratio (SJNR) is thus given by:

$$SJNR \approx 10\cdot\log_{10}\left(\frac{\frac{(k\cdot M)^2}{OSR\cdot f_{si}\cdot\sigma_j^2}}{(k\cdot M)^2\cdot\omega_x^2 + \frac{2}{3}\cdot\left\|H_{dNTF}\left(e^{i\omega}\right)\right\|_2^2}\right) \tag{10}$$

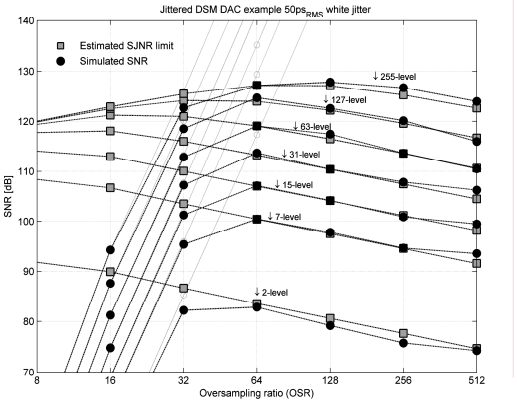


Figure 3 - SJNR estimate validation

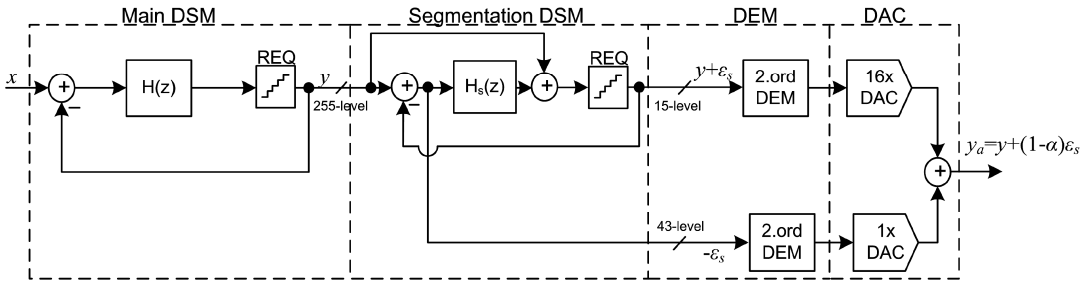


Figure 4 - 255-level DSM with segmented second-order mismatch shaping DAC

Typically k is between 0.5 and 0.8, higher with many levels and/or a conservative NTF. If M is very high, i.e. hi-res LPCM conversion, the first term dominates the denominator in (10) for large signals and the SJNR from the equation clearly converges towards that of regular sampling jitter. With few bits output the quantization noise will dominate and maximum SJNR increases 6dB per additional bit. Figure 3 shows validation of the SJNR estimate through Matlab simulations. The jitter is white PSD random jitter with σ_j=50ps. The DSM is very high order so jitter noise dominates quantization noise for OSR higher than 32.

It is apparent from fig.3 that using a high number of levels is very desirable. The problem in increasing M is DEM complexity and also in part DAC routing and thermometer encoding complexity. A proposed solution to this is to segment the DAC into two sub-DACs using a dedicated Segmentation-DSM (SDSM) to shape the inter sub-DAC mismatch [12]-[13]. This can be repeated in several steps for a higher degree of segmentation [14]. Figure 4 shows a 255-level DAC with one layer of shaped segmentation. In the presence of mismatch between the sub-DACs ($\alpha \neq 1$), the SDSM error ε_s leaks to the output and the SDSM NTF thus determines the inter sub-DAC mismatch shaping. A drawback with this approach is that the peak error from the SDSM is larger than the peak quantization error and therefore the lower sub-DAC needs additional levels. How many levels depend on the SDSM NTF; the number 43 comes from a second-order example SDSM also used in [13]. What this means is that the DAC in fig.4 has a combined range of 283·LSB rather than 255·LSB, implying some analog overhead. With several steps of segmentation, the overhead will be larger.

4. THE SEMIDIGITAL MULTI-BIT FIR DAC

An alternative to segmentation of the DAC using an SDSM is to revive the concept of the semidigital FIR-DAC, previously used to improve jitter performance with 1-bit DSM REQs [15]. Since the 1-bit DAC is inherently linear, parallel and time-delayed sub-DACs were in [15] used to obtain a multi-level output signal that did not suffer from nonlinearity. This can however also be done with a DEM-linearized multi-level DSM REQ driving multi-level sub-DACs connected as shown in fig.5.

All sub-DACs are here linearized with the same DEM block and mismatch between them will only lead to coefficient variations in the DAC's FIR filtering function. In-band the gain roughly equals the number of sub-DACs, so with 17 individual 15-level sub-DACs, the system is "255-level" at the output. Using a design example, we will show that the jitter performance will be comparable to a true 255-level DSM with a 255-level (non-segmented or segmented) DAC and that digital performance will not be a limiting factor.

The main reason why multi-bit DEM became prevalent in high-res audio instead of 1-bit modulators with semidigital FIR DACs, can probably be related to problems with stability, idle-tones and noise power modulation in the 1-bit DSM REQ itself [6],[16]. Historically it has proved difficult to eliminate these problems in 1-bit modulators, but as will be shown a multi-bit DSM REQ only needs quite few levels to render quantization noise and related artefacts negligible. The desire to increase the number of levels further is in an audio context primarily to improve wideband jitter immunity.

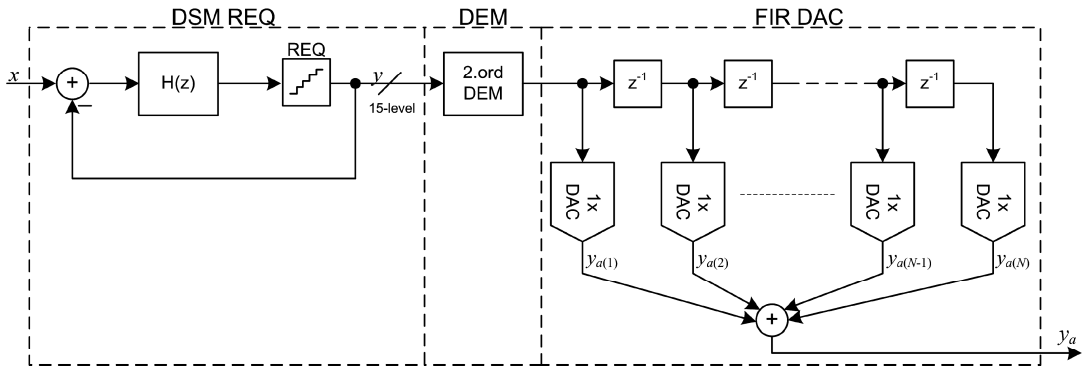


Figure 5 - Semidigital FIR-DAC driven by 15-level DSM REQ with second-order DEM

4.1. A design-example

In this section, a design-example is shown and results assessed under a given set of conditions. The theory presented is however general and other design parameters can be inserted into the equations at will. Comparisons are between three alternative DACs:

- A 15-level DSM with a 15-level DAC
- A 255-level DSM with 255-level segmented DAC
- A 15-level DSM with a 17·15-level FIR filter DAC

They all have the same NTF for the DSM:

$$H_{NTF}(z) = \frac{(z^2 - 2z + 1)(z^2 - 2z + 1)}{(z^2 - 0.8965z + 0.2198)(z^2 - 1.015z + 0.505)} \tag{11}$$

The oversampling ratio for the simulations shown is *OSR*=64. With the quantization noise approximation (7), maximum in-band signal-to-quantization noise ratio (SQNR) for the DSM REQ is calculated to be:

$$SQNR_{\max} \approx 10 \cdot \log_{10}\left(\frac{\frac{(k \cdot M)^2}{8}}{\frac{1}{12 \cdot 2\pi} \int_{\omega=-\frac{\pi}{OSR}}^{\frac{\pi}{OSR}} \left| H_{NTF}(e^{i\omega}) \right|^2 d\omega} \right) \tag{12}$$

The maximum stable k is found empirically and for this example is 0.8 or -2dBFS. In the 255-level case it's not limited by the DSM REQ itself, but rather by the SDSM which should not be overloaded [13]. $SQNR_{max}$ in the two cases is 152dB and 176dB, meaning that quantization noise is a negligible error source in a high-end audio context[1]. In the presence of 50ps RMS white jitter, $SJNR_{max}$ is estimated from (8)-(10) to be 107dB and 131dB respectively for the two cases. In other words it is confirmed that whereas quantization noise is negligible in both cases, *only the 255-level DAC* preserves genuine high-end audio performance in the presence of this modest amount of wideband clock jitter.

A semidigital FIR DAC as shown in fig.5 using N sub-DACs of equal weighting has the FIR function frequency response H_{DAC} being:

$$H_{DAC}(e^{i\omega}) = \frac{\sin\left(\frac{N\omega}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} \tag{13}$$

This is identical to the Dirichlet kernel or the aliased sinc function. Its DC or close to DC gain is approximately N. In-band there is a droop close to the Nyquist frequency that is small (less than a dB) as long as *OSR*>*N*. The droop can be compensated in the interpolation filter preceding the DAC and is disregarded in estimating the power of dy, now being:

[1] Typ. ~120dB total dynamic range.

$$\sigma_{dy}^2 \approx \frac{A^2}{8} \cdot \omega_x{}^2 \cdot N^2 + \frac{1}{12} \cdot \left\| H_{dNTF} \cdot H_{DAC} \right\|_2^2 \quad (14)$$

From (12) and using the NTF in (11), max SJNR is estimated to 132dB. The jitter performance is in other words at least as good as for a straightforward 255-level DAC, with only a single 15-level DEM circuitry.

4.2. Windowed Weighting

Weighting the sub-DACs equally might not be an optimal solution. In general it can be seen as a problem of minimizing the $H_{NTF} \cdot H_{DAC}$ L_2-norm. Therefore it is desirable to use a filter function with high damping in the region where the NTF is large. It should however also have coefficients implementable with reasonable hardware. An alternative that has proved to give good results for a wide variety of NTFs is weighting the sub-DACs according to a hann window [17]. The N-tap hann-window scaled for a DC-gain of N is given by the impulse response:

$$h_{DAC}[n] = 1 - \cos\left(2\pi \frac{n}{N-1}\right), 0 \le n \le N-1 \quad (15)$$

If the Dirichlet kernel of the N-tap rectangular window (13) is denoted as $D(\omega,N)$, the scaled hann window frequency response can be written as:

$$H_{DAC}(e^{i\omega}) = \left[D(\omega, N) + 0.5D\left(\omega - \frac{2\pi}{N-1}, N\right) + 0.5D\left(\omega + \frac{2\pi}{N-1}, N\right) \right] \cdot e^{i0.5\omega(N-1)} \quad (16)$$

Figure 6 shows the frequency response of the NTF (11) together with a rectangular 17-tap window as well as a 17-tap hann window. It is seen that the hann window has significantly better damping for high frequencies at the expense of a wider main lobe, but the NTF is still many dB down from its maximum gain across the entire main lobe.

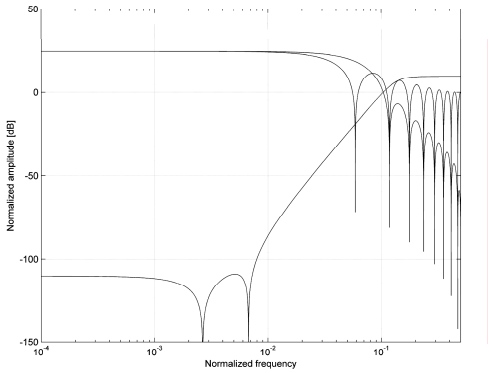


Figure 6 - Frequency response of NTF and of rectangular, hann window functions

For more aggressive NTFs, slightly better results may be achieved using a window function with an even wider main lobe and more attenuated side lobes, but simulations have shown hann windowing to be very versatile for most NTFs, and it is the window function of choice for the succeeding simulations. Inserting $H_{DAC}(e^{i\omega})$ from (16) into the SJNR calculations (8)-(10), a hann windowed semidigital DAC with 50ps RMS white jitter turns out to have an estimated maximum SJNR as high as 139dB.

5. JITTER PERFORMANCE SIMULATIONS

Figure 7 shows simulations of the output from a 15-level DSM REQ followed by a 15-level jittered DAC. There is no DAC mismatch. As seen the quantization noise from the DSM REQ is easily made negligible, with simulated SQNR at 152.4dB. However, when 50ps RMS white random jitter is introduced the SNR at the output degrades significantly, now being only 106dB and matching well with previous estimates.

The SJNR calculated from equations (8)-(10) is shown in comparison (theoretical SJNR), validating the estimation procedure and confirming jitter to be the limiting performance factor. The result shows how severely wideband jitter even in modest amounts will degrade the performance at the output of a typical 15-level DSM DAC.

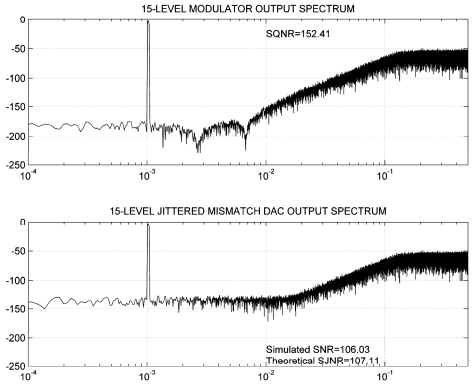


Figure 7 - 15-level DSM with 15-level jittered DAC, no mismatch

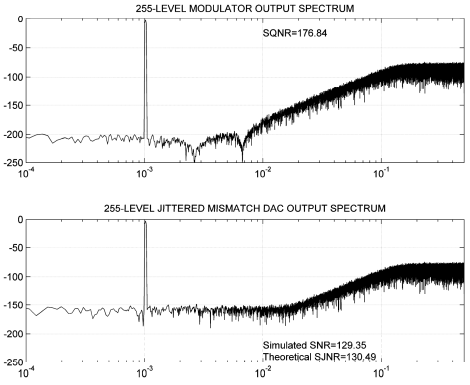


Figure 8 - 255-level DSM with 255-level jittered and segmented DAC, no mismatch

If the quantizer is increased from 15-levels to 255-levels (fig.8), using the same the DSM loop filter, the maximum SQNR as expected increases around 6dB per bit, from 152dB to almost 177dB. Introducing the same 50ps RMS white jitter now degrades the resolution to 129dB. This means the system clearly maintains high-end audio performance. The converter is implemented as a segmented DAC in accordance with the block diagram in fig.4.

In fig.9 the REQ is only 15-level, but a 255-element DAC is implemented as an equally weighted 17-tap semidigital FIR filter DAC. The simulated performance in the presence of the same jitter is about equal to the 255-level segmented DAC, as was also predicted in 4.1, and the maximum output SNR with 50ps RMS white jitter is now over 130dB.

For the simulations in fig.10, the DAC is implemented identically, but instead of equally weighted elements, sub-DACs are now weighted as a 17-tap hann-window. It is seen that the jitter performance now increases much more due to the superior out-of-band attenuation; the DAC having over 138dB SNR with 50ps RMS white jitter. This confirms that a semidigital FIR DAC with windowed weighting can improve jitter immunity quite substantially.

Conclusively, susceptibility to SNR deterioration from wideband jitter can be improved if a few-level DSM REQ is used in combination with a window weighted semidigital DAC. The DEM complexity remains low.

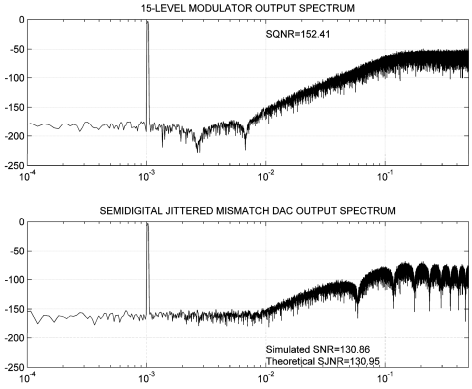


Figure 9 - 15-level DSM, 17-tap (255 elements total) semidigital FIR DAC, no mismatch

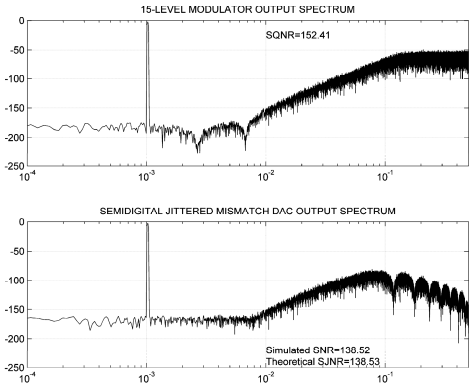


Figure 10 - 15-lev DSM, 17-tap (255 elements tot) hann-weighted semidigital FIR DAC, no mismatch

6. PERFORMANCE IN THE PRESENCE OF MISMATCH

Mismatch noise from a DAC with DEM can be estimated using an additive model for the mismatch error. The mismatch noise power is input dependent, but one can assume a worst-case scenario where $M/2$ elements are selected at any time (in other words mid-scale DAC input), and the mismatch error can then be approximated to be a white noise source with total power (see [9] for details):

$$\sigma_{\varepsilon_{mis}}^2 = \frac{\sigma_w^2 \cdot M}{4} \tag{17}$$

With an arbitrary mismatch-shaping function H_{DEM}, first- and second-order types having been shown in previous publications, the mismatch error is shaped into having a non-uniform PSD. The in-band signal-to-mismatch noise ratio (SMNR) of the DAC is found by integrating the error PSD over the Nyquist region:

$$SMNR_{\max} = \cdot 10 \cdot \log_{10} \left(\frac{\frac{(k \cdot M)^2}{8}}{\frac{\sigma_{\varepsilon_{mis}}^2}{2\pi} \int_{\omega=-\frac{\pi}{OSR}}^{\frac{\pi}{OSR}} \left| H_{DEM}(e^{i\omega}) \right|^2 d\omega} \right) \tag{18}$$

This means that if the element mismatch σ_w is constant relative to the quantizer step size, the maximum SMNR being proportional to M will increase by 3dB per bit. However since σ_w is typically inversely proportional to the physical element area [10], max SMNR will be constant if the *total* DAC area is kept constant while changing M. If a semidigital DAC is used, each sub-DAC is identically shaped. In the baseband the signal component from each sub-DAC adds linearly while the mismatch noise is assumed to be uncorrelated. In other words mismatch performance remains the same for the cases in fig.4 and fig.5 if the same total amount of elements is employed and the DAC die area is the same.

An additional problem in the semidigital case is that mismatch between sub-DACs leads to non-ideal coefficients in the DAC FIR filter function [18]. This compromises out-of-band noise suppression and consequently also reduces the DAC's jitter immunity.

Without loss of generality the semidigital DAC transfer function can be written as the sum of an ideal transfer function and an error transfer function:

$$H_{DAC}(e^{i\omega}) = H_{ideal}(e^{i\omega}) + H_{\varepsilon}(e^{i\omega}) \quad , \quad H_{\varepsilon}(e^{i\omega}) = \sum_{n=0}^{N-1} \varepsilon_{mis}[n] \cdot e^{i\omega n} \tag{19}$$

The mismatches in the sub-DACs are assumed to be real independent random variables, spectrally shaped by H_{DEM} and with a total (worst case) power given by (17). The expected PSD of the mismatch induced transfer function error in the filter DAC then becomes:

$$E\left\{ \left| H_{\varepsilon}\left(e^{i\omega}\right) \right|^2 \right\} = E\left\{ \sum_{k=0}^{N-1} \varepsilon_{mis}[k] \cdot e^{i\omega k} \cdot \sum_{l=0}^{N-1} \varepsilon_{mis}[l] \cdot e^{-i\omega l} \right\} \tag{20}$$

If a set of uncorrelated random variables are spectrally shaped by a filter function they are *still* uncorrelated, meaning that $E\{\varepsilon[k]\cdot\varepsilon[l]\}=0$ for $k \neq l$ and $E\{\varepsilon[k]\cdot\varepsilon[l]\}=\sigma_\varepsilon^2$ for $k=l$. In other words (20) will - regardless of DEM shaping - under the assumption of uncorrelated mismatch variables result in:

$$E\left\{ \left| H_{\varepsilon}\left(e^{i\omega}\right) \right|^2 \right\} = \sigma_{\varepsilon_{mis}}^2 \cdot \sum_{k=0}^{N-1} e^{i\omega k} \cdot e^{-i\omega k} = \sigma_{\varepsilon_{mis}}^2 \cdot N \tag{21}$$

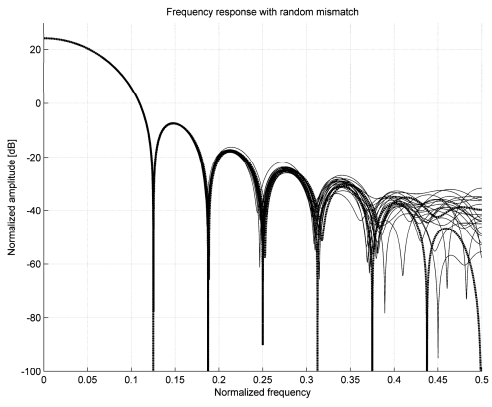


Figure 11 - Frequency response variation with sub-DAC mismatch

Since the expected amplitude response is given by the square root of the expected PSD, the estimated DAC filter function H_{DAC} with random element mismatch values is found to be:

$$E\left\{\left|H_{DAC}(e^{i\omega})\right|\right\}=\left|H_{ideal}(e^{i\omega})\right|+\sqrt{N}\cdot\sigma_{\varepsilon_{mis}} \quad (22)$$

The standard deviation of ε_{mis} is easily found as the square root of (17). Figure 11 shows the frequency response for 20 simulation runs of a 17-tap hann-weighted filter with 0.01·LSB RMS random coefficient mismatch. For this case the largest sub-DAC consists of 1.9·LSB elements and the smallest of 0.068·LSB elements. The ideal frequency response and the one estimated from (22) are shown in thick dashed lines. It is seen that the frequency response for different runs varies randomly around the estimated expectance value as they should. The variance will be determined by the number of coefficients and mismatch between them.

Confirming the analysis as valid for multi-bit DEM sub-DACs, fig.12 compares the output spectrum of a hann-weighted multi-bit FIR DAC with 1% RMS element mismatch and 2DWA mismatch-shaping, to filtering the output with a non-ideal filter given by (22). As is seen out-of-band noise suppression is compromised very similarly in the two cases (compare to fig.10), confirming (22) to be a good estimate for the frequency response of a multi-bit filter DAC with mismatch.

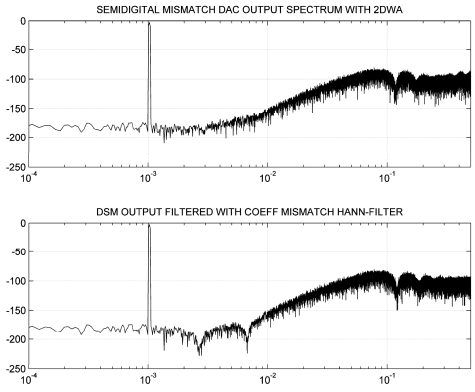


Figure 12 - Spectrum from real mismatch DAC with 2DWA DEM compared to (22)

As seen in fig.11, the effect on baseband response is negligible for all cases and it is the out-of-band damping of the filter DAC that is compromised. This as mentioned reduces jitter performance; how it affects the SJNR can be estimated by modifying (14), inserting (22) for $H_{DAC}(e^{i\omega})$:

$$\sigma_{dy}^2 \approx \frac{A^2}{8}\cdot\omega_x{}^2\cdot N^2 + \frac{1}{12}\cdot\left\|H_{dNTF}\cdot\left(H_{ideal}+\frac{\sqrt{N\cdot M}}{2}\cdot\sigma_w\right)\right\|_2^2 \quad (23)$$

In (23), N is the number of sub-DACs, M is the number of elements in each sub-DAC and σ_w is the RMS element mismatch normalized to the LSB.

7. JITTER AND MISMATCH PERFORMANCE SIMULATIONS

In this section the simulations from section 5 are repeated, now including DAC mismatch errors. Again a representative design example is used, but do note that the theory from both sections 4 and 6 is completely general. The DSM and DAC models are the same as in section 5, the jitter is still 50ps RMS wideband white jitter, but the DAC elements now have mismatch.

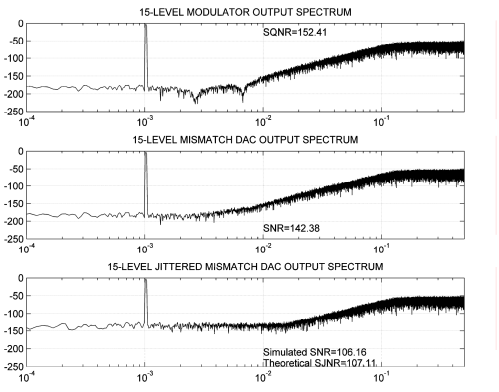


Figure 13 - 15-level DSM DAC performance with mismatch and white jitter

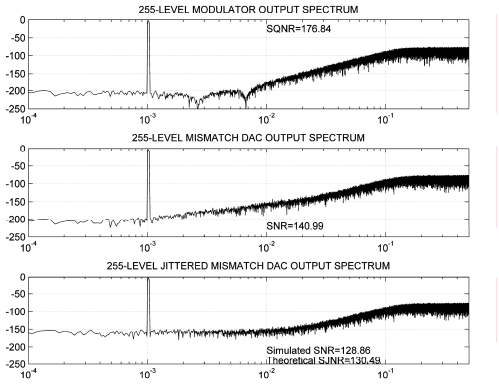


Figure 14 - 255-level segmented DAC performance with mismatch and white jitter

Element weights are Gaussian distributed random variables with a standard deviation of 1% at the 255-element LSB-level. The implemented DEM algorithm is a generic 2DWA [9], representative for second-order mismatch-shaping DEM that is typically used in high-end audio converters.

Figure 13 shows the output spectrum from a 15-level DAC. As is seen, mismatch with second order DWA only degrades the performance from 152dB to 142dB, but in the presence of 50ps RMS white jitter the resolution is jitter limited to 106dB just like it was without mismatch (fig.7). This DAC's performance is clearly jitter limited.

With a 255-level segmented DAC, shown in fig.14, the mismatch performance is somewhat lower. This is because the second order SDSM leakage caused by inter sub-DAC mismatch adds to the intra sub-DAC mismatch. Still it is clearly seen that the performance is jitter limited and approximately the same as without mismatch (fig.8). With 50ps RMS white jitter and 1% RMS mismatch, the segmented 255-level DSM DAC maintains high-end audio performance.

With a uniformly weighted semidigital DAC the mismatch performance should be around the same as for the 15-level DAC, which is confirmed by the middle sub-plot. The mismatch does however not compromise jitter performance which at more than 130dB is around the same as in fig.9. The damage mismatch infers on stopband damping does not affect the 17-tap rectangular filter notably, since it already has quite limited stopband damping as it is.

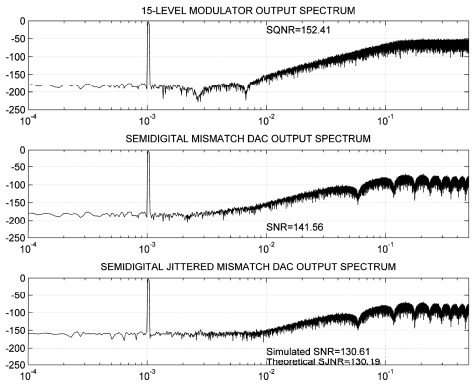


Figure 15 - Semidigital DAC performance with mismatch and white jitter

The 17-tap hann-weighted semidigital FIR DAC in fig.16 on the other hand has visibly compromised stopband damping when compared to fig.10. Nonetheless the damping - and from it the SJNR - is clearly superior to both the uniform FIR DAC and the segmented 255-level DAC. The SMNR does affect the final SNR more though, which is not unexpected knowing the sub-DACs are differently weighted and the smallest coefficient only 1/20 of the largest. As seen from the figure, the SMNR and SJNR contribute about equally to the SNR reduction, and together they bring the total resolution down to around 135dB. This is still a significant performance improvement over the uniformly weighted DAC. It is beneficiary to have different error sources contributing at a similar level since it indicates nothing is over-designed.

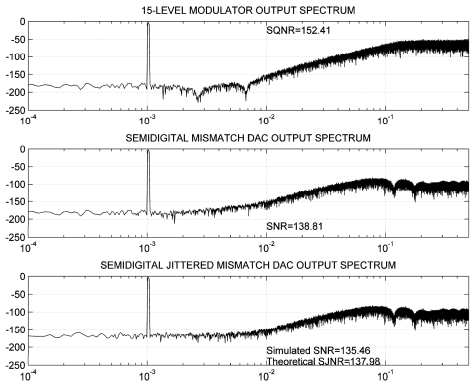


Figure 16 - Hann-weighted semidigital DAC performance with mismatch and white jitter

With a different mismatch standard deviation or with a different clock jitter magnitude, the error budget could become different and perhaps not be in the hann-weighted DACs favour. Especially if there is significantly less anticipated jitter (than 50ps RMS) and higher mismatch (than 1% RMS at the 8-bit LSB-level), the error budget will less favourable.

From practical experience, it is however more likely that mismatch will be *lower* and the jitter levels *higher* than the numbers used in this example. If that is the case the hann-weighted semidigital DAC will be *more* advantageous compared to the alternatives.

8. CONCLUSIONS

This paper shows how to make estimates for jitter performance in continuous time delta-sigma based audio DACs. Using many bits in the DAC can improve immunity to wideband clock jitter significantly. It is shown that in the presence of 50ps RMS white jitter a high number of bits are needed to maintain high-end audio performance. To use many bits in the DSM and DAC implies high complexity for the DEM switching network. This must be alleviated, which has previously been proposed done with segmented DEM [12]-[14]. This paper shows that reduced complexity DEM and better jitter performance can be achieved by instead implementing a semidigital multi-bit FIR filter DAC. For a typical audio OSR, 64 used in the simulation examples in this paper, few bits are needed to render quantization noise negligible. Arranging extra DAC elements as a weighted filter DAC can be more beneficiary than using extra bits in the DSM REQ. Theory to estimate performance reduction from intra and inter sub-DAC mismatch is derived. Simulations confirm that with 50ps RMS white jitter and 1% mismatch at the 8-bit LSB level, a hann-weighted filter DAC performs approximately 6dB better than a segmented DSM REQ and DAC with the same analog complexity. With higher jitter levels and/or lower mismatch levels, it will be even more beneficiary.

9. ACKNOWLEDGEMENTS

This work was supported by the Norwegian Research Council under grant 162101 SPECK.

10. APPENDIX

The impact of mismatch on the DAC frequency response in (22) differs from the result in [18], where a very complicated mathematical treatment led to a white error expression with a slightly (1dB) lower PSD. It is the author's opinion that (22) is correct and its proof is given in this appendix.

A *completely* general expression for the error filter PSD is given by:

$$\left|H_\varepsilon\left(e^{\mathrm{i}\omega}\right)\right|^2 = \sum_{k-0}^{N-1}\varepsilon_{mis}[k]\cdot e^{\mathrm{i}\omega k}\cdot\sum_{l-0}^{N-1}\varepsilon_{mis}^{*}[l]\cdot e^{-\mathrm{i}\omega l} \tag{24}$$

This expression can be expanded as follows:

$$\begin{aligned}\left|H_\varepsilon\left(e^{\mathrm{i}\omega}\right)\right|^2 = &\underbrace{\sum_{k=0}^{N-1}\varepsilon_{mis}[k]\cdot\varepsilon_{mis}^{*}[k]}_{k=l}\\ &+\underbrace{\sum_{k=1}^{N-1}e^{-\mathrm{i}\omega k}\sum_{l=k}^{N-1}\varepsilon_{mis}[l]\cdot\varepsilon_{mis}^{*}[l-k]}_{k>l}\\ &+\underbrace{\sum_{k=1}^{N-1}e^{\mathrm{i}\omega k}\sum_{l=k}^{N-1}\varepsilon_{mis}^{*}[l]\cdot\varepsilon_{mis}[l-k]}_{k<l}\end{aligned} \tag{25}$$

Using Euler's formula, (25) is simplified to:

$$\begin{aligned}\left|H_\varepsilon\left(e^{\mathrm{i}\omega}\right)\right|^2 = &\sum_{k=0}^{N-1}\left|\varepsilon_{mis}^{2}[k]\right|^2\\ &+2\cdot\sum_{k=1}^{N-1}\cos(\omega k)\cdot\sum_{l=k}^{N-1}\mathrm{Re}\left\{\varepsilon_{mis}[l]\cdot\varepsilon_{mis}[l-k]\right\}\end{aligned} \tag{26}$$

The power spectrum derived in (26) is completely general. All mismatch values are of course real numbers and they are also, as pointed out in the main text, assumed to be uncorrelated Gaussian random variables. This means ε_{mis} has the basic autocorrelation function:

$$E\left\{\varepsilon_{mis}[l]\varepsilon_{mis}[l-k]\right\} = \begin{cases}\sigma_{\varepsilon_{mis}}^2 & ,k=l\\ 0 & ,\text{otherwise}\end{cases} \tag{27}$$

It follows from this that the expected error filter power spectrum, found by inserting (27) in (26), then becomes:

$$E\left\{\left|H_{\varepsilon}\left(e^{i\omega}\right)\right|^{2}\right\}=N\cdot\sigma_{\varepsilon_{mis}}^{2} \quad (28)$$

And the expected amplitude response as the root of the power spectrum will thus be the constant:

$$E\left\{\left|H_{\varepsilon}\left(e^{i\omega}\right)\right|\right\}=\sqrt{N}\cdot\sigma_{\varepsilon_{mis}} \quad (29)$$

Note also that (26) can be used to calculate the error response exactly for any deterministic set of mismatch values.

11. REFERENCES

[1] M.O.J.Hawksford: "Jitter Simulation in High Resolution Digital Audio", *Audio Engineering Society Convention Paper 6864*, 121st Convention San Francisco, October 2006

[2] P.Heydari; "Analysis of the PLL Jitter Due to Power/Ground and Substrate Noise", *IEEE Trans. Circuits and Systems Part-I: Regular Papers,* vol.51, no.12, pp.2404–2416, Dec.2004

[3] C.Dunn, M.O.J.Hawksford; "Is the AESEBU/ SPDIF Digital Audio Interface Flawed?", *Audio Engineering Society Convention Paper 3360*, 93rd Convention, San Francisco, October 1992

[4] J.Dunn; "Jitter: Specification and Assessment in Digital Audio Equipment", *Audio Engineering Society Convention Paper 3361*, 93rd Convention, San Francisco, October 1992

[5] AES-12id-2006; AES information document for digital audio measurements—Jitter performance specifications

[6] R.Schreier; "Understanding Delta-Sigma Data Converters", *IEEE Press*, ISBN 0-471-46585-2, 2005

[7] R.W.Adams; "Design and Implementation of an Audio 18-Bit Analog-to-Digital Converter Using Oversampling Techniques", *J. Audio Eng. Soc*., vol. 34 No.3, pp. 153-166, March 1986

[8] I.Fujimori, A.Nogi, T.Sugimoto; "A Multibit Delta-Sigma Audio DAC With 120-dB Dynamic Range", *IEEE J. Solid-State Circuits*, vol.35, no.8, pp.1066-1073, Aug.2000

[9] R.K.Henderson and O.Nys; "Dynamic Element Matching Techniques with Arbitrary Noise Shaping Function", *Proc. IEEE Int. Symp. Circuits and Systems ISCAS'96*, pp.293-296, May 1996

[10] M.J.M. Pelgrom et al., "Matching properties of MOS transistors", *IEEE J. Solid-State Circuits*, 24(5), pp. 1433-1439, October 1989

[11] H.T.Jensen and I.Galton, "A low-complexity dynamic element matching DAC for direct digital synthesis," *IEEE Trans. of Circuits and Systems Part-II*, vol. 45.1, pp. 13-27, Jan. 1998

[12] R.Adams, K.Nguyen, K. Sweetland, "A 112-dB SNR Oversampling DAC with Segmented Noise Shaped Scrambling", *Audio Engineering Society Convention Paper 4774*, 105th Convention San Fransisco, Sept.1998

[13] I.Løkken, A.Vinje, T.Sæther, "Segmented Dynamic Element Matching using Delta-Sigma Modulation, *Proc. AES 31st International Conference,* London, UK, June 2007

[14] J.Steensgaard-Madsen; "High Performance Data Converters", *Ph.D. thesis, Technical University of Denmark Dept. Information Technology*, March 1999

[15] D. Su and B. Wooley, "A CMOS Oversampling D/A Converter With a Current-Mode Semi-Digital Reconstruction Filter," *IEEE J. Solid-State Circuits*, vol. 28, pp. 1224-1233, Dec. 1993

[16] J.Vanderkooy, S.Lipshitz, "Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications", *Audio Engineering Society Convention Paper 5395*, 110th Convention Amsterdam, May 2001

[17] R.B.Blackman and J.W.Tukey: "Particular Pairs of Windows," in *The Measurement of Power Spectra, From the Point of View of Communications Engineering.* New York: Dover, 1959

[18] A.Petraglia and S.K.Mitra: "Effects of Coefficient Inaccuracy in Switched-Capacitor Transversal Filters", *IEEE Trans. Circuits and Systems*, vol.38, no.9, pp.977-983, Sept.1991.

Bibliography

[1] T.A.Edison, “Phonograph or Speaking Machine”, US. Patent 200,521, (1878 Feb.)

[2] H.Fletcher and W.A.Munson, “Loudness, its Definition, Measurement and Calculation”, *J. Acoust. Soc. Am*., vol.5, pp.82-108, (1933 May)

[3] D.W.Robinson and R.S.Dadson, "A Re-Determination of the Equal-Loudness Relations for Pure Tones", *Br. J. Appl.Phys*., vol.7, pp.166-181, (1956)

[4] T.Oohashi et al., “Inaudible High-Frequency Sounds Affect Brain Activity: Hypersonic Effect”, *J.Neurophysiol*, vol.83 no.6, pp.3548-3558, (2000 Jun.)

[5] S.Kiryu, “Detection of Threshold for Tones Above 22kHz”, *Audio Eng. Soc. Convention Paper 5401*, 110th AES Convention, Amsterdam, (2001 May)

[6] J.Boyk, “There’s Life Above 20kHz! – A Survey of Musical Instrument Spectra to 102.4kHz”, *California Institute of Technology*, available on-line at: http://www.cco.caltech.edu/~boyk/spectra/spectra.htm, (2000 May)

[7] Acoustic Renaissance for Audio, “A Proposal for High-Quality Application of High-Density CD Carriers”, *J. Japan Audio Soc.,* vol.35, available on-line at: http://www.meridian-audio.com/ara, (1995 Oct.)

[8] J.Atkinson, A.B.Krueger, “The Great Debate; Subjectivism on Trial”, *Home Entertainment Show 2005*, New York. Audio recording available on-line at: http://stereophile.com/news/050905debate/, (2005 May)

[9] G.E.Moore, “Cramming More Components Onto Integrated Circuits”, *Electronics Magazine*, vol.38, no.8, (1965)

[10] Philips Intellectual Property and Standards, “IEC-908: Compact Disc Digital Audio – The Red Book”, *International Electrotechnical Commission Standards Document*, no. 28/10/04-3122 783 0027 2, (1980 Jun.)

[11] G.Theile, “On the Performance of Two-Channel and Multi-Channel Stereophony”, *Audio Eng. Soc. Convention Paper 2887*, 88th AES Convention, Montreux, (1990 Mar.)

[12] P.Craven, “Toward the 24-bit DAC: Novel Noise-Shaping Topologies Incorporating Correction for the Nonlinearity in a PWM Output Stage”, *J. Audio Eng. Soc.*, vol.41, no.5, pp.291-313, (1993 May)

[13] J.v.d.Verbakel, L.v.d.Kerkhof, M.Maeda, Y.Inazawa, ” Super Audio CD Format”, *Audio Eng. Soc. Convention Paper 4705,* 104th AES Convention, Amsterdam (1998 May)

[14] N.Fuchigami, T.Kuroiwa, B.H.Suzuki, “DVD-Audio Specifications”, *J. Audio Eng. Soc.*, vol.48, no.12, pp.1228-1230, 1232-1238, 1240; (2000 Dec.)

[15] D.Blech, M.C.Yang, “DVD-Audio Versus SACD: Perceptual Discrimination of Digital Audio Coding Formats”, *Audio Eng. Soc. Convention Paper 6086,* 116th AES Convention, Berlin, (2004 May)

[16] J.Atkinson, “Hi-Rez Media: When Will They Learn?”, *Stereophile Magazine*, vol.28, no.3, (2005 Mar.)

[17] P.J.Alexander, “Peer-to-Peer File Sharing: The Case of the Music Recording Industry”, *Review of Industrial Organization*, vol.20, pp.151-161, (2002)

[18] T.Painter, A.Spanias, “Perceptual Coding of Digital Audio”, *Proc. IEEE*, vol.88, no.4, pp.451-515, (2000 Apr.)

[19] F.E.Toole, “Loudspeaker Measurements and Their Relationship to Listener Preferences: Part I-II”, *J. Audio Eng. Soc*., vol.34, no.4, pp. 227-235 and no.5, pp.323-348, (1986)

[20] R.Levine, “The Death of High Fidelity”, *Rolling Stone Magazine*, available on-line at: http://www.rollingstone.com/news/story/17777619, (2007 Dec.)

[21] H.Nyquist, “Certain Topics in Telegraph Transmission Theory", *Trans. AIEE*, vol.47, pp.617-644, (1928 Apr.)

[22] V.A.Kotelnikov, “On the Carrying Capacity of the Ether and Wire in Telecommunications”, *1st All-Union Conference on Questions of Communication*, Lzd. Red. Upr. Svyazi RKKA, Moscow, (1933)

[23] C.E.Shannon, “A Mathematical Theory of Communication”, *Bell System Tech. J*., vol.27, pp.379–423, 623–656, (1948)

[24] C.E.Shannon, "Communication in the Presence of Noise", *Proc. Institute Radio Eng*., vol.37, no.1, pp.10-21, (1949 Jan.)

[25] E.T.Whittaker, “On the Functions Which are Represented by the Expansions of the Interpolation Theory”, *Proc. Royal Soc. Edinburgh*, vol.35, pp.181-194, (1915)

[26] H.D.Lüke, “The Origins of the Sampling Theorem”, *IEEE Communications Magazine,* pp.106–108, (1999 Apr.)

[27] H.S.Black and J.O.Edson, "Pulse Code Modulation," *AIEE Transactions***,** vol.66, pp.895-899 (1947)

[28] W.R.Bennett, "Spectra of Quantized Signals," *Bell System Tech. J.*, vol.27, pp.446-471, (1948 July)

[29] B.Widrow; “A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory", *IRE Trans. Circuit Theory*, vol.CT-3, pp.266-276, (1956 Dec.)

[30] R.M.Gray, “Quantization Noise Spectra,” *IEEE Trans. Inform. Theory,* vol.36, (1990 Nov.)

[31] G.R.Ritchie, J.C.Candy, and W.H.Ninke, "Interpolative Digital-to-Analog Converters", *IEEE Trans. Communications*, vol.22, pp.1797-1806, (1974 Nov.)

[32] H.G.Musmann and W.W.Korte, "Generalized Interpolative Method for Digital/Analog Conversion of PCM Signals", U.S. Patent 4,467,316**,** (filed 1981 June)

[33] M.Bellanger et al., ”Digital Filtering by Polyphase Network: Application to Sample-Rate Alteration and Filter Banks”, *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.24, pp.109-114, (1976 Apr.)

[34] T.Saramäki, “Design of FIR Filters as a Tapped Cascaded Interconnection of Identical Subfilters,” *IEEE Trans. Circuits and Systems,* vol.34, no.9, pp.1011-1029, (1987 Sept.)

[35] T.Saramäki, Y. Neuvo, S.K.Mitra, “Design of Computationally Efficient Interpolated FIR-filters,” *IEEE Trans. Circuits and Systems*, vol.35, no.1, pp.70-88, (1988 Jan.)

[36] O.Pirochta, ”Hardware Implementations of Digital FIR Filters in FPGA”, *Proc.17th International Conference Radioelektronika 2007*, Brno Czech Republic, (2007 Apr.)

[37] S.Mitra, “Digital Signal Processing: A Computer-Based Approach”, third edition, *McGraw-Hill International Press,* ISBN: 007-124467-0, (2006)

[38] L.G.Roberts, "Picture Coding Using Pseudo-Random Noise". *IEEE Trans. Information Theory*, vol.8, pp.145–154, (1962 Feb.)

[39] L.Schuchman, "Dither Signals and Their Effect on Quantization Noise", *IEEE Trans. Communications*, vol.12, pp.162–165 (1964 Dec.)

[40] S.P.Lipshitz, J.Vanderkooy, “Dither in Digital Audio”, *J. Audio Eng.Soc.*, vol.35, no.12, pp.966-975, (1987 Dec.)

[41] S.P.Lipshitz, R.A.Wannamaker, and J.Vanderkooy, “Quantization and Dither; A Theoretical Survey,” *J. Audio Eng. Soc.,* vol. 40, pp. 355–375 (1992 May).

[42] S.P.Lipshitz, J.Vanderkooy, “Dither Myths and Facts”, *Audio Eng. Soc. Convention paper 6279*, 117th AES Convention, San Francisco, (2004 Oct.)

[43] R.A.Wannamaker, “The Theory of Dithered Quantization”, Ph.D. Thesis, Dept. for Applied Mathematics, University of Waterloo, Waterloo, ON, Canada (1997 June)

[44] H.Inose, Y.Yasuda and J.Marakami; “A Telemetering System by Code Modulation, Delta-Sigma Modulation”, *IRE Trans. Space, Electronics and Telemetry*, SET-8, pp. 204-209, (1962 Sept.)

[45] D.J.Goodman, “The Application of Delta Modulation of Analog-to-PCM Encoding”, *Bell System Tech.J.*, vol.48, pp.321-343, (1969 Feb.)

[46] C.C.Cutler, “Transmission Systems Employing Quantization”, U.S. Patent 2,927,962, (1960 Mar.)

[47] Dan Sheingold, “Sigma-Delta or Delta-Sigma?”, *Analog Dialogue*, vol.24, no.2, editors’ note, (1990)

[48] N.H.C.Gilchrist, ”Analogue-to-Digital and Digital-to-Analogue Converters for High Quality Sound”, *Audio Eng. Soc. Convention Paper 1583*, 65th AES Convention, London, (1980 Feb.)

[49] R.W.Adams; “Design and Implementation of an Audio 18-Bit Analog-to-Digital Converter Using Oversampling Techniques”, *J. Audio Eng. Soc.*, vol.34 no.3, pp.153-166, (1986 March).

[50] J.T.Caves, M.A.Copeland, C.F.Rahim and S.D.Rosenbaum, “Sampled-Data Filters Using Switched Capacitors as Resistor Equivalents”, *IEEE J. Solid-State Circuits*, vol.12, pp.592-600, (1977 Dec.)

[51] J.A.C.Bingham, “Application of Direct-Transfer SC Integrator,” *IEEE Trans. Circuits and Systems,* vol.31, pp.419–420, (1984 Apr.).

[52] N.S.Sooch et al., “18-b Stereo D/A Converter with Integrated Digital and Analog Filters”, *Audio Eng. Soc. Convention Paper 5603*, 91st AES Convention, New Your, (1991 Oct.)

[53] I.Fujimori, A.Nogi, T.Sugimoto; **“**A Multibit Delta-Sigma Audio DAC With 120-dB Dynamic Range*”, IEEE J. Solid-State Circuits*, vol.35, no.8, pp.1066-1073 (2000 Aug.)

[54] I.Fujimori, T.Sugimoto, “A 1.5 V, 4.1 mW Dual-Channel Audio Delta–Sigma D/A Converter”, *IEEE J. Solid-State Circuits*, vol.33, no.12, pp.1863-1870, (1998 Dec.)

[55] A.Paul Brokaw, “Digital-to-Analog Converter with Current Source Transistors Operated Accurately at Different Current Densities," U.S. Patent 3,940,760, (filed 1975 Mar.)

[56] N.Terada, S.Nakao, “A 126DB D-Range Current-Mode Advanced Segmented DAC”, *Proc. 16th Audio Eng. Soc. UK Conference – Silicon for Audio*, (2001 Mar.)

[57] K.B.Amulya, “Binomial Theorem in Ancient India”, *Indian J. Hist. Sci.*, vol.1, pp.68–74, (1966)

[58] G.Boole, “An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities”, Macmillan Publishers (1854), reprinted with corrections, Dover Publications, New York, ISBN 978-0486600284, (1958).

[59] C.Shannon, "The Symbolic Analysis of Relay and Switching Circuits", *Trans. Am. Inst. Electrical Eng.*, vol.57, pp.713-723, (1938 Mar.).

[60] J.J.Wikner, “Studies on CMOS Digital-to-Analog Converters”*,* Ph.D. Thesis, Dept. for Electrical Engineering, Linköping University, Linköping, Sweden, ISBN 91-7219-910-5, (2001)

[61] Q.Li, “INL, DNL and Performance of Analog-to-Digital Converters”, project report for the course Learning from Data at Portland State university, available on-line at: http://web.cecs.pdx.edu/~edam/Reports/2002/Li.pdf

[62] C.Dunn, M.O.J.Hawksford; "Is the AESEBU/SPDIF Digital Audio Interface Flawed?", *Audio Eng. Soc. Convention Paper 3360*, 93rd AES Convention, San Francisco, (1992 Oct.)

[63] J.Dunn; "Jitter: Specification and Assessment in Digital Audio Equipment", *Audio Eng. Soc. Convention Paper 3361*, 93rd AES Convention, San Francisco, (1992 Oct.)

[64] J.Dunn et al., "Toward Common Specifications for Digital Audio Interface Jitter", *Audio Eng. Soc. Convention Paper 3705*, 95th AES Convention, New York, (1993 Oct.)

[65] J.Dunn, "Sample Clock Jitter and Real-time Audio Over the IEEE1394 High Performance Serial Bus", *Audio Eng. Soc. Convention Paper 4920*, 106th AES Convention, Munich, (1999 Apr.)

[66] P.Heydari; **"**Analysis of the PLL Jitter Due to Power/Ground and Substrate Noise*"*, *IEEE Trans. Circuits and Systems I: Regular Papers,* vol.51, no.12, pp.2404–2416, (2004 Dec.)

[67] J.A.McNeill; "Jitter in Ring Oscillators", *IEEE J. Solid State Circuits*, vol.32, no.6, pp.870-879, (1997 June)

[68] AES-12id-2006; AES Information Document for Digital Audio Measurements — Jitter Performance Specifications, (2006)

[69] M.O.J.Hawksford: "Jitter Simulation in High Resolution Digital Audio", *Audio Eng. Soc. Convention Paper 6864*, 121st AES Convention San Francisco, (2006 Oct.)

[70] K.Doris, A.van Roermund, D. Leenaerts; "A General Analysis on the Timing Jitter in D/A Converters", *Proc. IEEE Intern. Symp. Circuits and Systems ISCAS 2002*, pp.117-120, (2002 May)

[71] L.Angrisani, M.D'Apuzzo, M.D'Arco; "Modeling Timing Jitter Effects in Digital-to-Analog Converters", *2005 IEEE International Workshop on Intelligent Signal Processing*, pp.254-259, (2005 Sept.)

[72] B.Putzeys, R.de saint Moulin, "Effects of Jitter on AD/DA Conversion", *Audio Eng. Soc.Convention Paper 6122*,, 116th AES Convention, Berlin, (2004 May)

[73] R.H.M. van Veldhoven; "A Triple-Mode Continuous-Time ΣΔ Modulator with Switched-Capacitor Feedback DAC for a GSM-EDGE/CDMA2000/UMTS receiver", *IEEE J. Solid State Circuits*, vol.38, no.12, pp.2069-2076, (2003 Dec.)

[74] K.Ashihara et al., "Detection Threshold for Distortions Due to Jitter on Digital Audio", *ACJ J. Acoust. Science and Technology*, vol.26, no.1, pp.50-54 (2005)

[75] R.W.Adams, "Jitter Analysis of Asynchronous Sample-Rate Conversion", *Audio Eng. Soc. Convention Paper 3712*, 95th AES Convention New York, (1993 Oct.)

[76] F.M.Rotacher; "Sample-Rate Conversion; Algorithms and VLSI Implementation", PhD-thesis, Swiss Federal Institute of Technology, Zürich, (1995)

[77] M.J.M. Pelgrom et al., "Matching Properties of MOS Transistors", *IEEE J. Solid-State Circuits*, vol.24, no.5, pp.1433-1439, (1989 Oct.).

[78] K.O.Andersson, J.J.Wikner; "Modeling of the Influence of Graded Element Matching Errors in CMOS Current-Steering DACs", *Proc. 17th Norchip Conference*, Oslo Norway, (1999 Nov.).

[79] M.Clara, A.Wiesbauer,W.Klatzer; "Nonlinear Distortion in Current-Steering D/A-Converters Due to Asymmetrical Switching", *Proc. IEEE Intern. Symp. Circuits and Systems ISCAS 2004*, pp.285-288, (2004 May).

[80] B.P.Del Signore et al.; "A Monolithic 20-b Delta-Sigma A/D Converter", *IEEE J. Solid-State Circuits*, vol.25, no.6, pp.1311-1317, (1990 Dec.)

[81] Luschas S. and Lee H.-S., "Output Impedance Requirements for DACs", *Proc. IEEE Intern. Symp. Circuits and Systems ISCAS 2003*, pp.861–864, (2003 May)

[82] Texas Instruments, "DSD1792A 24-Bit 192 kHz Sampling Advanced Segment Audio Stereo DAC", Data Sheet SLES106 Rev.B, (2006 Nov.)

[83] J.Silva, U.Moon, J.Steensgaard and G.C.Temes "Wideband Low-Distortion Delta-Sigma ADC Topology", *Electronic Letters*, vol.37, no.12, pp.737-738, (2001 June)

[84] J.C.Candy, "A Use of Double Integration in Sigma-Delta Modulation", *IEEE Trans. Communications*, vol.33, no.3, pp.249-258, (1985 Mar.)

[85] R.W.Adams, "Theory and Practical Implementation of a Fifth-Order Sigma-Delta A/D Converter", *J. Audio Eng. Soc.*, vol.39, no.7/8, pp.515-528, (1991 Jul./Aug.)

[86] B.E.Boser, B.A.Wooley, "The Design of Sigma-Delta Modulation Analog-to-Digital Converters", *IEEE J. Solid State Circuits*, vol.23, no.6, pp.1298-1308, (1988 Dec.)

[87] S.Hein, A.Zakhor, "On the Stability of Sigma Delta Modulators", *IEEE Trans. Signal Processing*, vol.41, no.7, pp.2322-2348, (1993 Jul.)

[88] S.Lipshitz, J.Vanderkooy, R.A.Wannamaker, "Minimally Audible Noise Shaping", *J. Audio Eng. Soc*., vol.39, no.11, pp.836-852, (1991 Nov.)

[89] H.Takahashi, A.Nishio "Investigation of Practical 1-bit Delta–Sigma Conversion for Professional Audio Applications", *Audio Eng. Soc. Convention Paper 5392*, 110th AES Convention, Amsterdam, (2001 Apr.)

[90] P.J.Naus, et al., "A CMOS Stereo 16-bit D/A Converter for Digital Audio", *IEEE J. Solid State Circuits*, vol.22, no.3, pp.390-395, (1987 June).

[91] P.Kiss, J.Arias, D.Li, V.Boccuzzi, "Stable High-Order Delta-Sigma DACs", *IEEE Trans. Circuits and Systems I, Reg.Papers*, vol.51, no.1, pp.200-205, (2004 Jan.)

[92] T. Hayashi et al., "A Multistage Delta-Sigma Modulator without Double Integration Loop", *ISSCC Dig. Technical Papers*, pp.182-183, (1986 Feb.)

[93] Y.Matsuya et al., "A 16-bit Oversampling A-to-D Conversion Technology using Triple-Integration Noise Shaping", *IEEE J. Solid State Circuits*, vol.22, no.6, pp.921-929, (1987 Dec.)

[94] W.Chou et al., "Multistage Sigma-Delta Modulation", *IEEE Trans. Information Theory*, vol.35, no 4, (1989 July)

[95] H.Kato, "Trellis Noise-Shaping Converters and 1-bit Digital Audio", *AES Convention Paper 5615*, 112th AES Convention, Munich Germany, (2002 May).

[96] A.J.Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. Information Theory*, vol.13, no.2, pp.260–269, (1967 Apr.)

[97] E.Janssen, D.Reefman, "Advances in Trellis based SDM structures", *AES Convention Paper 5993*, 115th AES Convention, New York USA, (2003 Oct.)

[98] J.A.S.Angus, "The Efficiency of "Pruned Tree" versus "Stack" Algorithms for Look-Ahead Sigma-Delta Modulators", *J. Audio Eng. Soc*., vol.54, no.6, pp.477-494, (2006 June)

[99] W.L.Lee, C.G.Sodini, "A Toplogy for Higher-Order Interpolative Coders", *Proc. IEEE Int. Symp. Circuits and Systems*, pp.459-462, (1987 May)

[100] D.L.Wellard et al., "Delta-Sigma Modulator with Oscillation Detect and Reset Circuit", US Patent 5,012,244, (1991 Apr.)

[101] W.Rhee, B.S.Song, A.Ali, "A 1.1-GHz CMOS Fractional-N Frequency Synthesizer with a 3-b Third-Order ΔΣ Modulator", *IEEE J. Solid State Circuits – Part I*, vol.35, no.10, pp.1453-1460, (2000 Oct.)

[102] E.F.Stikvoort, "Some Remarks on the Stability and Performance of the Noise Shaper or Sigma-Delta Modulator," *IEEE Trans. Comm.,* vol.36, pp.1157-1162, (1988 Oct.)

[103] T.Ritoniemi, T.Karema, H.Tenhunen; "Design of Stable High Order 1-bit Sigma-Delta Modulators", *Proc. 1990 IEEE Int. Symp. Circuits and Systems*, pp.3267–3270, (1990 May)

[104] R.Schreier; "An Empirical Study of High-Order Single Bit Delta-Sigma Modulators", *IEEE Trans. Circuits and Systems – Part II*, vol.40, pp.461-466, (1993 Aug.)

[105] S.H.Ardalan,J.J.Paulos, "Analysis of Nonlinear Behaviour in Delta-Sigma Modulators", *IEEE Trans. Circuits and Systems – Part I*, vol.34, no.6, pp.593-603, (1987 June)

[106] S.Hein, A.Zakhor, "On the Stability of Sigma Delta Modulators," *IEEE Trans. Signal Processing*, vol.41, no.7, pp.2322-2348, (1993 July)

[107] H.Wang, **"**A Study of Sigma Delta Modulations as Dynamical Systems," PhD Thesis, Colombia University, New York, AAT 9333879, (1993)

[108] O.Feely, L.O.Chua, “The Effect of Integrator Leak in ΣΔ Modulation”, *IEEE Trans. Circuits and Systems*, vol.38, no.11, pp.1293-1305, (1991 Nov.)

[109] L.Risbo, “Delta-Sigma Modulators: Stability Analysis and Optimization”, PhD. thesis, Technical University of Denmark, Lyngby, Denmark, (1994 June)

[110] J.Reiss; “Towards a Procedure for Stability Analysis of High Order Sigma Delta Modulators”, *Audio Eng. Soc. Convention Paper 6549*, 119th AES Convention, New York, (2005 Oct.)

[111] G.Tsenov, V.Mladenov, and J.Reiss, “A Comparison of Theoretical, Simulated, and Experimental Results Concerning the Stability of Sigma Delta Modulators, *Audio Eng. Soc. Convention Paper 7440*, 124th AES Convention, Amsterdam, (2008 May)

[112] M.Goodson, B.Zhang, and R.Schreier, "Proving Stability of Delta-Sigma Modulator Using Invariant Sets," *Proc. Int. Symp Circuits and Systems ISCAS’95*, pp.633-636, (1995 May)

[113] R.Schreier, M.Goodson, B.Zhang, "An Algorithm for Computing Convex Positively Invariant Sets for Delta-Sigma Modulators," *IEEE Trans. Circuits and Systems Part-I: Fundamental Theory and Applications*, vol. 44, no.1 pp.38-44, (1997 Jan.)

[114] J.C.Candy, O.J.Benjamin, “The Structure of Quantization Noise from Sigma-Delta Modulation”, *IEEE Tran. Communications*, vol.29, no.9, pp.1316-1323 (1981 Sept.)

[115] V.Friedman, “The Structure of the Limit Cycles in Sigma Delta Modulation”, *IEEE Trans. Communications*, vol.36, no.8, pp.972-979 (1988 Aug.).

[116] D.Reefman, J.Reiss, E.Janssen, M.Sandler, “Description of Limit Cycles in Sigma-Delta Modulators”, *IEEE Trans. Circuits and Systems I; Regular Papers*, vol.52, no.6, pp.1211-1223, (2005 June)

[117] J.Reiss, M.Sandler, “They Exist: Limit Cycles in High Order Sigma Delta Modulators”, *Audio Eng. Soc. Convention Paper 5832*, 114th AES Convention, Amsterdam, (2003 Feb.)

[118] J.Reiss, “Understanding Sigma-Delta Modulation: The Solved and Unsolved Issues”, *J. Audio Eng. Soc.*, vol.56, no.1/2, pp.49-64, (2008 Jan.)

[119] R.Schreier, “On the Use of Chaos to Reduce Idle-Channel Tones in Delta-Sigma Modulators”, *IEEE. Trans. Circuits and Systems Part-I*, vol.41, no.8, pp.539-547 (1994 Aug.)

[120] S.R.Norsworthy, “Dynamic Dithering of Delta-Sigma Modulators”, *Audio Eng. Soc. Convention Paper 4103*, 99th AES Convention, New York, (1995 Oct.)

[121] J.Reiss, M.Sandler, “Dither and Noise Modulation in Sigma-Delta Modulators”, *Audio Eng. Soc. Convention Paper 5935*, 115th AES Convention, New York, (2003 Oct.)

[122] D.Campbell, “The Delta-Sigma Modulator as a Chaotic Dynamical Non-Linear System”, PhD. thesis, University of Waterloo, Ontario, Canada, (2006)

[123] J.G.Kenney and L.R.Carley, "Design of Multibit Noise-Shaping Data Converters", *J. Analog Int. Circuits Signal Processing*, vol.3, no.3, pp.259-272, (1993 May)

[124] R.J.Van De Plassche; "Dynamic Element Matching for High Accuracy Monolithic DA Converters", *IEEE J. Solid State Circuits*, vol.11, no.6, pp.795-800, (1976 Dec.)

[125] L.R.Carley, "A Noise Shaping Coder Topology for 15+ bit Converters", *IEEE J. Solid State Circuits*, vol.24, no.2, pp.267-273, (1989 Apr.)

[126] B.H.Leung and S.Sutarja, "Multibit Sigma-Delta A/D Converter Incorporating a Novel Class of Dynamic Element Matching Techniques", *IEEE Trans. Circuits and Systems Part-II: Analog and Digital Signal Processing*, vol.39, no.1, pp.35-51, (1992 Jan.)

[127] R.T.Baird, T.S.Fiez, "Linearity Enhancement of Multi-Bit Δ-Σ A/D and D/A Converters Using Data Weighted Averaging", *IEEE Trans. Circuits and Systems Part-II: Analog and Digital Signal Processing*, vol.42, no.12, pp.753-762, (1995 Dec.)

[128] M.Vadipour, "Techniques for Preventing Tonal Behaviour of Data Weighted Averaging Algorithm in ΔΣ-Modulators", *IEEE Trans. Circuits and Systems Part-II*, vol.47, no.11, pp 1137-1144, (2000 Nov.)

[129] A.A.Hamoui and K.Martin, "Linearity Enhancement of Multibit ΔΣ Modulators Using Pseudo Data-Weighted Averaging", *Proc. IEEE International Symp. Circuits and Systems ISCAS'02*, pp.III 285-288, (2002 May)

[130] K.D.Chen and T.H.Kuo, "An Improved Technique for Reducing Baseband Tones in Sigma-Delta Employing Data Weighted Averaging Algorithms without Adding Dither", *IEEE Trans. Circuits and Systems Part-II*, vol.46, no.1, pp 53-68, (1999 Jan.)

[131] R.K.Henderson and O.Nys, "Dynamic Element Matching Techniques with Arbitrary Noise Shaping Function", *Proc. IEEE Int. Symp. Circuits and Systems ISCAS'96*, pp.293-296, (1996 May)

[132] X.M.Gong, **"**An Efficient Second-Order Dynamic Element Matching Technique for a 120 dB Multi-Bit Delta-Sigma DAC", *Audio Eng. Soc. Convention Paper 5124*, 108th AES Convention, Paris, (2000 Feb.)

[133] R.W.Adams, "Data Directed Scrambler for Multi-Bit Noise Shaping D/A Converters", U.S.Patent no. 5,404,142, (1995 Apr.)

[134] I.Galton, "Spectral Shaping of Circuit Errors in Digital-to-Analog Converters," *IEEE Trans. Circuits and Systems Part-II: Analog and Digital Signal Processing,* vol. 44, no. 10, pp. 808-817, (1997 Nov.)

[135] J.Welz, I.Galton, E.Fogleman, "Simplified Logic for First-Order and Second-Order Mismatch-Shaping Digital-to-Analog Converters," *IEEE Trans. Circuits and Systems Part-II: Analog and Digital Signal Processing*, vol.48, no.11, pp.1014-1028, (2001 Nov.)

[136] E.Fogleman, J.Welz, I.Galton, "An Audio ADC Delta-Sigma Modulator with 100-dB Peak SINAD and 102-dB DR Using a Second-Order Mismatch-Shaping DAC," *IEEE J. Solid-State Circuits*, vol. 36, no. 3, p.339-348, (2001 Mar.)

[137] E.N. Aghdam, P. Benabes, "Higher Order Dynamic Element Matching by Shortened Tree-Structure in Delta-Sigma Modulators", *Proceedings of the 2005 European Conference Circuit Theory and Design*, vol.1, pp.I/201- I/204, (2005 Sept.)

[138] R.Schreier, B.Zhang "Noise-Shaped Multibit D/A Converter Employing Unit Elements," *Electronic Letters*, vol.31, no.20, pp.1712-1713, (1995 Sept.).

[139] J.A.Schoeff, "An Inherently Monotonic 12 Bit DAC," *IEEE J. Solid State Circuits,* vol.14, no.6, pp.904-911, (1979 Dec.)

[140] R.Adams, K.Nguyen, K.Sweetland, "A 112dB SNR Oversampling DAC with Segmented Noise Shaped Scrambling", *Audio Eng. Soc. Convention Paper 4774*, 105th AES Convention, San Francisco, (1998 Sept.)

[141] J.Steensgaard-Madsen, "High Performance Data Converters", Ph.D. thesis, Technical University of Denmark Dept. Inf. Tech., (1999)

[142] A.Fishov, E.Siragusa, J.Welz, E.Fogleman, I.Galton, "Segmented Mismatch-Shaping D/A Conversion", *Proc. IEEE International Symp. Circuits and Systems ISCAS'02*, (2002 May)

[143] M.O.J.Hawksford, "Digital-to-Analog Converter with Low Intersample Transition Distortion and Low Sensitivity to Sample Jitter and Transresistance Amplifier Slew Rate," *J. Audio Eng.Soc.,* vol.42, no. 11, pp.901-917, (1994 Nov.).

[144] R.Adams, K.Nguyen, K.Sweetland, "A 113dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling", *IEEE J. Solid State Circuits*, vol.33, no.12, pp.1871-1878, (1999 Dec.)

[145] M.Clara, W.Klatzer, A.Wiesbauer, D.Straeussnigg; "A 350MHz Low-OSR Delta-Sigma Current-Steering DAC with Active Termination in 0.13µm CMOS", *Proc. IEEE International Solid-State Circuits Conference ISSCC 2005*, pp.118-588, (2005 Feb.)

[146] D. Su and B. Wooley, "A CMOS Oversampling D/A Converter with a Current-Mode Semi-Digital Reconstruction Filter," *IEEE J. Solid-State Circuits*, vol. 28, pp. 1224-1233, (1993 Dec.)

[147] W.R.Bennett, ""New Results in the Calculation of Modulation Products", *Bell System Tech. J.*, vol. 12, pp. 228-243, (1933 Apr.)

[148] B.D. Josephson, "Pulse Width Modulated Audio Amplifiers", *Wireless World*, letter to the editor, vol.71, pp. 335-336, (1965 July).

[149] J.D.Martin**,** "Theoretical Efficiencies of Class-D Power Amplifiers", *Proc. IEE.***,** vol.117, no.6, pp.1089-1090, (1970)

[150] Y. Mitsuhashi, "Mathematical Analysis of a Pulse Width Modulation Digital to Analog Converter", *J. Audio Eng.Soc*., vol.31, no.3 pp.135-138; (1983 Mar.)

[151] M.Sandler, "Towards a Digital Power Amplifier", *AES Convention Paper 2135*, 76th AES Convention, New York, (1984 Oct.)

[152] List of PWM amplifiers, http://www.avsforum.com/avs-vb/showthread.php?t=594707

[153] A.Hewitt, "A Simple Approximation for the Distortion in a Pulse-Width-Modulation Digital-to-Analogue Converter", *Audio Eng.Soc. Convention Paper 4598*, 103rd AES Convention, New York (1997 Sept).

[154] J.M.Goldberg, M.B.Sandler, "Noise Shaping and Pulse-Width Modulation for an All-Digital Audio Amplifier, *J.Audio Eng.Soc*., vol.39, no.6, pp.449-460, (1991 June)

[155] K.Nielsen, "Audio Power Amplifiers with Energy Efficient Power Conversion", Ph.D. Thesis, Tech.University of Denmark, Lyngby, (1998 Apr.)

[156] E.Gaalaas, B.Y.Liu, N.Nishimura, R.Adams, K.Sweetland, "Integrated Stereo ΔΣ Class-D Amplifier", *IEEE J. Solid State Circuits*, vol.40, no.12, (2005 Dec.).

[157] R.Khoini-Poorfard, D.A.Johns, "On the Effect of Comparator Hysteresis in Interpolative ΔΣ Modulators", *Proc. Int. Symp. Circuits and Systems ISCAS'93*, vol.2, pp.1148-1151, (1993 May)

[158] T.S.Doorn, E.van Tuijl et al., "An Audio FIR-DAC in a BCD Process for High-Power Class-D Amplifiers", *Proc. 31st European Solid State Circuits Conference ESSCIRC 2005*, pp.459-462, (2005 Sept.)

[159] T.Rueger et al., "A 110dB Ternary PWM Current-Mode Audio DAC with Monolithic 2Vrms Driver", *Int. Solid State Circuits Conference Digest of Tech. Papers ISSCC 2004*, vol.1, pp.372-533, (2004 Feb.)

[160] D.Reefman, J.v.d.Homberg, E.van Tuijl, et al., "A New Digital to-Analog Converter Design Technique for HiFi Applications," *Audio Eng. Soc. Convention Paper 5846*, 114th AES Convention, Amsterdam, (2003 March)

[161] P.Kiss, J.Arias, D.Li, and V.Boccussi, "Stable High-Order Delta-Sigma DACs", *IEEE Trans. Circuits and Systems Part I: Regular Papers*, vol.51, no.1, (2004 Jan.)

[162] Y.Cheng, C.Petrie, B.Nordick, D.Comer, "Multibit Delta-Sigma Modulator with Two-Step Quantization and Segmented DAC", *IEEE Trans. Circuits and Systems Part II: Express Briefs*, vol.53, no.9, pp.848-852, (2006 Sept.)

[163] H.J.Schouwenaars et al., "A Monolithic Dual 16-bit D/A Converter", *IEEE J. Solid State Cicuits*, vol.21, no.3, pp.424-429, (1986 Jun.)

[164] P.J.A.Naus et al., "A CMOS Stereo 16-bit D/A Converter for Digital Audio", *IEEE J. Solid State Circuits*, vol.22, no.3, pp.390-395, (1987 Jun.)

[165] J.Sneep et al., "A Bit-Stream Digital to Analog Converter with 18-b Resolution", *IEEE J. Solid State Circuits*, pp.1757-1763, vol.26, no.12, (1991 Dec.)

[166] S.Nakano et al., "A 117dB D-Range Current-Mode Multi-Bit Audio DAC for PCM and DSD Audio Playback", *Audio Eng. Soc. Convention paper 5190*, 109th AES Convention, Los Angeles, (2000 Sept.)

[167] R.H.Walden, "Analog-to-Digital Converter Survey and Analysis", *IEEE J. Selected Areas in Communications*, vol.17, pp.539-550, (1999 Apr.)

[168] AES17-1998 (r2004): AES Standard Method for Digital Audio Engineering – Measurement of Digital Audio Equipment (Revision of AES17-1991, (1998)

[169] P.Duhamel, M.Vetterli, "Fast Fourier Transforms: A Tutorial Review and State of the Art", *J. Signal Processing*, vol.19, no.4, pp.259-299 (1990 Apr.)

[170] J.W.Gibbs, "Fourier Series", *Nature* 59, 200 (1898) and 606 (1899).

[171] R.B.Blackman and J.W.Tukey: "Particular Pairs of Windows", *The Measurement of Power Spectra, From the Point of View of Communications Engineering.* New York: Dover, (1959)

[172] F.J.Harris, "On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform", *Proc. of the IEEE*, vol.66, no.1, pp.51-83, (1978 Jan.)

[173] Maxim Application Note 1040, "Coherent Sampling vs. Window Sampling", available on-line from: http://www.maxim-ic.com/an1040

[174] J.Blair, "Histogram Measurement of ADC Nonlinearities Using Sine Waves", *IEEE Trans. Instrumentation and Measurement*, vol.43, pp.373-383, (1994 June)

[175] I.Løkken, A.Vinje, "Some Considerations for Spectral Analysis of Delta-Sigma Data Converters", *ISAST Trans. Electronics and Signal Processing*, accepted for publication