Fourier Analysis and Synthesis for a Mildly Non-Linear Quantizer

A.J.E.M. Janssen[1]
Philips Research Laboratories Eindhoven, WY-81
5656 AA Eindhoven, The Netherlands
e-mail: A.J.E.M.Janssen@philips.com

&

N. Csizmadia
Philips Semiconductors Nijmegen, M3010
6534 AE Nijmegen, The Netherlands
e-mail: Nico.Csizmadia@philips.com

[1]Corresponding author.

Number of pages: 31

Number of figures: 13

Keywords: Quantizer, THD, SINAD, Non-Linearity, Fourier Analysis, Fourier Synthesis, Wobbling, Dithering, AD-Converter.

Abstract.
We introduce and analyze a method for stabilizing test parameters and estimating the non-linearity of mildly non-linear quantizers by Fourier analysis of the quantizer's output to sinusoidal input signals. Stabilization for the combined effects of the rounding operation of the quantizer and the variations in offset and amplitude of the applied sinusoid is achieved by a deterministic form of subtractive dithering, called wobbling. It thus appears that one can retrieve low-to-medium degree non-linearities whose contribution to the quantizer's distortion in the relevant frequency range is of the same order as or well below the distortion due to rounding before wobbling. For the analysis of the method a thorough investigation of the spectra of sampled, quantized sinusoids (wobbled or not) has to be made. As such we complete and extend existing results in the literature of spectra of this type. The method is potentially useful for estimating the integral non-linearity of AD-converters.

Fourier Analysis and Synthesis for a
Mildly Non-Linear Quantizer

A.J.E.M. Janssen, N. Csizmadia

1. Introduction

We start by explaining the background of the problem that we intend to solve in this paper. In the testing practice of Analog-to-Digital (AD-) converters there are several test parameters aimed at assessing the quality of the converter with respect to its global and local non-linear behaviour, the noise due to its internal circuitry, the absence of codes, jitter, etc. Because test times are to be kept to a minimum, one is particularly interested in obtaining the test parameters by applying sinusoidal input signals so that the outputs can be analyzed using FFTs. Direct time methods often need time-averaging because of the noise and are thus much more time-consuming in nature. Also, the demands on the accuracy of the test signals in these time methods (such as being near-to-perfectly linear over a large range) can be hard to fulfill. Not all sources of distortions and imperfections of an AD-converter can be located using test parameters obtained from frequency-domain measurements, but it is very well conceivable that the global non-linearity is one of the features that can be assessed via the Fourier domain. Normally one should expect such a global non-linearity to be determined by only a small number of higher harmonics of the AD-converter's output to a sinusoidal input. In that case the influence of the noise can be largely ignored (or might even be helpful) when estimating the global non-linearity.

In this paper we elaborate the above point of view. One of the major problems one is faced with in this approach is the distortion at the harmonics due to the rounding operation of the AD-converter. It turns out that this distortion strongly varies with the variations in offset and amplitude of the input sinusoid. Variations of less than 0.1 quantization step, which are quite common in the testing practice, may result in variations of the rounding distortion of more than 10 dB. Since it is common AD-converter's design practice to aim at converters where the distortions due to the rounding and to the global non-linearity are of the same order of magnitude, one cannot hope for a reliable estimation of the (distortion due to the) non-linearity unless certain measures are taken to suppress the rounding distortion.

We introduce and elaborate in this paper a method, called wobbling, in which the distortion due to rounding when using sinusoidal input signals is substantially reduced. We analyze and assess its performance under the ideal circumstances that the AD-converter has a perfect quantizer, no jitter, no missing codes and that the influence of the noise can be ignored for most of our considerations. That is, we assume an input x – output y relation of the AD-converter of the form

(1.1) $$y = [h(x) + \mathbf{n}]\,,$$

where h is a relatively smooth, continuous-valued function (global non-linearity), $\mathbf{n}$ is Gaussian white noise (device noise) of zero expectation and variance σ^2 with σ well below the quantization step size, and [ ] represents the rounding operation with unit quantization step (quantizer). See [2], [10] where a similar model is used. The [ ] refers to ideal rounding-to-the-nearest-integer and is given by

$$[x] = \underset{n\in\mathbb{Z}}{\arg}\ \min |x-n| \ , \qquad x \in \mathbb{R} \ , \tag{1.2}$$

with the convention that $[n - \frac{1}{2}] = n$ for integer n. It is evident that the model used here is far too simple to describe all phenomena that take place in real AD-converters; for instance, the fact that in real AD-converters the distance between the decision levels of the quantizers are non-uniform is completely ignored. Rather, the model (1.1) describes the input-output relation of what one may call a mildly non-linear, somewhat noisy quantizer. We shall not try to define "mildly non-linear" in a precise mathematical sense; one should think of a global non-linearity h, with $h(x)-x$, $h'(x)-1$ small on a large range of $x \in \mathbb{R}$, that can be approximated well by a polynomial of not too high degree. The advantage of using such a simple model is that one can mathematically analyze the performance of the wobble method in dependence of its parameters so that one can come to recommendations when such a method is to be used for stabilizing test parameters and calibration for global non-linearities of practical AD-converters.

2. Frequency-domain test parameters

We shall now introduce the frequency-domain test parameters THD^2 and SINAD^2 that are commonly used in the testing practice for assessing the quality of non-linear noisy devices. We consider input signals

$$x(t) = A_0 \sin 2\pi M t - c_0 \ , \qquad 0 \le t < 1 \ , \tag{2.1}$$

where M is a positive integer and the amplitude A_0 and offset c_0 are close to their nominal values A_n and 0, with A_n normally equal or somewhat less than $\frac{1}{2}$ times the full range of the device. When $y(t)$ is the device's output on the input $x(t)$, we let

$$Y(\nu) = \int_0^1 e^{-2\pi i\nu t}\, y(t)\, dt \ , \qquad \nu \in \mathbb{R} \ , \tag{2.2}$$

be the Fourier transform of $y(t)$. Then we define

$$\mathrm{THD}^2 = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |Y(nM)|^2 \ , \tag{2.3}$$

where L is an integer of the order 5–10, and

(2.4) $$\text{SINAD}^{-2} = \frac{1}{\frac{1}{2}A_0^2} \int_0^1 |y(t) - x(t)|^2 \, dt \, .$$

Actually, the A_0 occurring at the right-hand sides of (2.3–4) might be unknown, in which case we assume that a reliable and fast estimate of it can be obtained from the output $y(t)$. Also, see Sec. 7 (in particular (7.5)) for this point. In (2.3) THD stands for Total Harmonic Distortion and SINAD stands for SIgnal-to-Noise-And-Distortion ratio. Under the model (1.1) the THD is thought as being due to the rounding operation of the quantizer and the global non-linearity h, where the latter should be a polynomial of degree $\leq L$. Also, the SINAD is supposed to reflect the combined effects of the (rounding and global) non-linearities and the noise.

In both the THD^2 and SINAD^2 the influence of the rounding operation is prominently present. To assess this influence, we consider the case of a perfect quantizer ($h(x) = x$, noise $\mathbf{n} = 0$ in (1.1)). It will be shown in Sec. 5 that in good approximation (when A_0 is not too small, say $A_0 \geq 5$)

(2.5) $$\text{THD}^2 \approx \frac{8(L-1)}{\pi^2 A_0^3} \{\zeta^2_{-\frac{1}{2}}(A_0 - c_0 - \tfrac{1}{2}) + \zeta^2_{-\frac{1}{2}}(A_0 + c_0 - \tfrac{1}{2})\} \, ,$$

(2.6) $$\text{SINAD}^{-2} \approx \frac{2}{A_0^2} \Big\{ \frac{1}{12} - \frac{4\sqrt{2}}{3\pi\sqrt{A_0}} (\zeta_{-\frac{3}{2}}(A_0 - c_0 - \tfrac{1}{2}) + \zeta_{\frac{3}{2}}(A_0 + c_0 - \tfrac{1}{2})) \Big\} \, ,$$

where $\zeta_{-\frac{1}{2}}$ and $\zeta_{-\frac{3}{2}}$ are particular 1-periodic functions depicted in Fig. 1. Hence the THD^2 shows considerable fluctuations when A_0, c_0 vary, and the deviation of SINAD^{-2} from $1/6A_0^2$ is also quite noticeable. The $\frac{1}{12}$ in the right-hand side of (2.6) between $\{ \ \}$ is the familiar number for the average quantization error power with unit stepsize as occurs in many places in literature.

In Sec. 9 we shall present a formula for reconstructing a global non-linearity h, assumed to be a polynomial of degree $\leq L$, on the range $|x| \leq A_0$ exactly from the ideal Fourier coefficients

(2.7) $$Y_{\text{id}}(nM) = \int_0^1 h(A_0 \sin 2\pi M t) \, e^{-2\pi i n M t} \, dt \, , \qquad |n| \leq L \, .$$

It is clear that, given the fact that the distortion due to the non-linearity h and the rounding are of the same order of magnitude (the latter showing large fluctuations), one cannot expect the formula in Sec. 9 to yield reliable reconstruction results when the $Y_{\text{id}}(nM)$ of (2.7) are replaced by the $Y(nM)$ as obtained from the output $y(t)$, see (2.2).

3. Stabilizing frequency-domain test parameters: wobbling

We now present a method for stabilizing the lower harmonics $Y(nM)$, see (2.2), and the SINAD, see (2.4), for the combined effect of the rounding operation and small variations of

A_0 and c_0 in the input $x(t)$ in (2.1). A well-known method in audio engineering to reduce the effect of the rounding operation, at least perceptually, is dithering [9], [6], [3], [8]. Here one adds noise $\mathbf{n}_d$ to the input signal $x(t)$ with a probability density function (pdf) such that the average value after rounding equals $x(t)$. To avoid higher order distortion, manifesting itself audibly, one subtracts this $\mathbf{n}_d$ again from the output (subtractive dither), or, when this is not feasible, one makes special choices for the pdf so that at least the first and second moment of the output are equal to the corresponding moments of the input. Because of timing problems between input and output (a misalignment of one sample distance already being disastrous), subtractive dither is often not an option; this is, for instance, so in the testing practice of AD-converters. When non-subtractive dither is used, it turns out, see Sec. 7, that the SINAD^{-2} gets an offset, so that it becomes a problem how to properly ascribe the measured SINAD^{-2} values to the rounding, device noise and dither noise. Furthermore, when sampling rates are kept as low as possible (as one prefers to do in AD-converter testing), the finiteness of the sample sizes seriously draws upon the performance of dithering methods.

The setting of the problem we deal with in this paper is much more of a deterministic nature than the one encountered in audio engineering. While the variety of input signals in audio engineering is enormous, we have sinusoidal inputs of a given frequency and of amplitudes and offsets that vary in a very limited range. It is, therefore, natural to try to use a form of dithering that is more deterministic in nature as well. We propose to do this as follows. We replace, for each t in the periodicity interval $[0, M^{-1})$ of the input $x(t)$ and each $p = 0, ..., M-1$, the input $x(t + p/M) = x(t)$ by

$$x\Big(t + \frac{p}{M}\Big) + g\Big(t + \frac{p}{M}\Big)\ , \tag{3.1}$$

and we subtract $g(t + p/M)$ again from the output on (3.1). Here $g(t)$, $0 \le t < 1$, is a deterministic function such that for any $t \in [0, M^{-1})$ the set of numbers

$$\Big\{g\Big(t + \frac{p}{M}\Big)(\mathrm{mod}\,\mathbb{Z})\ |\ p = 0, ..., M-1\Big\} \tag{3.2}$$

is uniformly distributed in $[0, 1)$. One thus achieves that the mean values

$$\frac{1}{M}\sum_{p=0}^{M-1}\Big(\Big[x\Big(t + \frac{p}{M}\Big) + g\Big(t + \frac{p}{M}\Big)\Big] - g\Big(t + \frac{p}{M}\Big)\Big) \tag{3.3}$$

are, on average, as close to $x(t)$ as possible, so that the harmonics of the signal $[x(t) + g(t)] - g(t)$ can be expected to be as close to those of $x(t)$ as possible.

The natural examples of signals $g(t)$, $0 \le t < 1$ that meet the uniform distribution condition of the sets (3.2) for $0 \le t < M^{-1}$ are ramps

$$g(t) = r_Q(t) = Q(t - \tfrac{1}{2})\ , \qquad 0 \le t < 1\ , \tag{3.4}$$

with integer span Q such that $\gcd(Q, M) = 1$. The problem of misalignment of input $x(t) + g(t)$ and output $[x(t) + g(t)]$ due to imperfect timing is not severe when Q is not too large and timing errors are small. One can achieve very good alignment between input and output signal by realizing that both signals should have very dominant Fourier components at frequencies $\pm M$ with phases $\mp \pi/2$, so that an adjustment suggests itself on comparing the phases of the Fourier components at frequencies $\pm M$.

The simplest ramp of the form (3.4) has $Q = 1$, and we shall call this ramp the standard ramp. Nevertheless, the ramps with higher values of Q are of practical significance. Taking a higher Q, one can achieve a better code visitation of the input while a certain robustness against non-uniformity of the quantizer's decision levels as occurs in practical AD-converters is obtained. See [4], Sec. 5.

To sum up we propose the following method which we succintly refer to as wobbling. The term "wobbling" is also used in other applications, such as the addressing and timing control in recordable CDs and DVD (here sinusoidal wobble signals are used). We take as input

$$(3.5) \qquad x(t) = A_0 \sin 2\pi M t + r_Q(t) \ , \qquad 0 \le t < 1 \ ,$$

and we compute the harmonics

$$(3.6) \qquad Y^{sw}(nM) = \int_0^1 y^{sw}(t)\, e^{-2\pi i n M t} \ , \qquad n \in \mathbb{Z} \ ,$$

of the subtractively wobbled output signal

$$(3.7) \qquad y^{sw}(t) = y(t) - r_Q(t) \ , \qquad 0 \le t < 1 \ ,$$

with $y(t)$ the output to $x(t)$. These $Y^{sw}(nM)$ are used in the definition (2.3) of THD^2 to replace $Y(nM)$ as well as in the formula in Sec. 9 to estimate the non-linearity h. Furthermore, we replace in the definition (2.4) of SINAD^{-2} the input $x(t)$ of (2.1) by $x(t)$ of (3.5) and the output $y(t)$ by $y^{sw}(t)$ of (3.7).

We shall analyze this method under the assumption that the model (1.1) is valid with a noise $\mathbf{n}$ that can be ignored when the THD^2 is considered and that has sufficiently small correlation time when the SINAD^{-2} is considered. Also, we assume that we may linearize the non-linearity h so that

$$(3.8) \qquad h(A_0 \sin 2\pi M t + r_Q(t)) = h(A_0 \sin 2\pi M t) + r_Q(t) \ .$$

For validity of the latter assumption, one should restrict to non-linearities h that are sufficiently mild and to ramps r_Q whose span Q is modest compared to A_0.

We have not tried the proposed method for use in audio engineering to counteract the perceptive annoying effects of quantization. There are reasons to doubt usefulness of the method in this area, since all effort has been spent to have the method work for sinusoidal input signals of a given frequency only. When investigating this point further, it may be useful to try other wobble signals g than the ramp (for instance, triangular functions) that meet the condition of uniform distribution of the sets in (3.2).

4. Overview of results

We now present an overview of the results obtained in this paper. In [7], which is available on request or, alternatively, can be downloaded from

http://www.research.philips.com/manuscript/search.html ,

a much more detailed exposition is given. We shall often refer to [7] for additional information, related results, proofs, etc., but the present paper is sufficiently self-contained that availability of [7] is not strictly required.

In Sec. 5 we consider the Fourier spectrum of the quantization error

$$e(t) = [A_0 \sin 2\pi Mt - c_0] - (A_0 \sin 2\pi Mt - c_0) = \sum_{n=-\infty}^{\infty} b_n \, e^{2\pi i n Mt} \; . \tag{4.1}$$

The obtained results serve as a reference for what follows since they deal with the non-wobbled case with a perfect quantizer ($h(x) = x$, no noise). We thus present Bessel series expressions for the b_n's, their asymptotics as $A_0 \to \infty$ and n stays bounded and the consequences for the THD^2. We also give results on the global shape of the spectrum of $e(t)$; these are relevant to the developments in Sec. 8 where the influence of sampling is considered. We furthermore present a Bessel series representation and its asymptotics as $A_0 \to \infty$ for the average quantization error power

$$\int_0^1 e^2(t) \, dt = \sum_{n=-\infty}^{\infty} |b_n|^2 \; . \tag{4.2}$$

Many of the results just discussed can be extended to more general error signals

$$[f(t)] - f(t) \; ; \qquad f(t) = A \sin 2\pi Mt - c(t) \tag{4.3}$$

with $f(t)$, $c(t)$ periodic of period M^{-1} and $c(t)$ smooth and of relatively small amplitude. The prime example of an f as in (4.3) is

$$f(t) = h(A_0 \sin 2\pi Mt - c_0) \tag{4.4}$$

with h a mild non-linearity. For the latter type of signals we indicate that the THD^2 is in good approximation additive with respect to the distortion due to the rounding only and due to the non-linearity only.

In Sec. 6 we consider M^{-1}-periodic signals f as above subtractively wobbled according to

$$f^{sw}(t) = [f(t) + g(t)] - g(t) \; . \tag{4.5}$$

We discuss conditions for best wobbling, and we indicate that signals $g(t)$, $0 \le t < 1$, satisfying the uniform value distribution condition, see (3.2), are optimal in this respect. We present the counterparts of the results given in Sec. 5 for sinusoids that are wobbled using a ramp. That is, we consider the harmonics b^{sw}_{nM} of the subtractively wobbled error signal

$$(4.6)\qquad e^{sw}(t) = [A_0 \sin 2\pi Mt - c_0 + r_Q(t)] - (A_0 \sin 2\pi Mt - c_0 + r_Q(t)) = \\ = \sum_{k=-\infty}^{\infty} b^{sw}_k \, e^{2\pi ikt} \; .$$

As a consequence we find that, for the THD^2 and SINAD^{-2}, the effect of wobbling using a ramp of span Q with $\gcd(Q, M) = 1$ amounts qualitatively to replacing the amplitude A_0 and offset c_0 by MA_0 and Mc_0, respectively. Alternatively, one may say that for the THD^2 and SINAD^{-2} wobbling using $r_Q(t)$ amounts qualitatively to adding ${}^2\log M$ bits to the quantizer.

In Sec. 7 we consider the SINAD^{-2} for the case of a non-perfect quantizer ($h(x) \neq x$; noise present). It is indicated that, when wobbling is used or not used,

$$(4.7)\qquad \mathrm{SINAD}^{-2} = \frac{1}{\frac{1}{2}A_0^2} \left(\tfrac{1}{12} + \sigma^2 + S_h^2 + \Delta\right) ,$$

where σ^2 is the variance of the noise, S_h^2 is a distortion term closely related to the THD^2 due to the non-linearity h only, and Δ is an error term tending to 0 as $A_0 \to \infty$ and which is substantially smaller in the case that wobbling is used. The fact that the right-hand side of (4.7) comprises the various sources of noise and distortion incorporated in the model (1.1) in an additive way is very satisfying. We also make a connection with some basic results in the theory of non-subtractive dithering as found in [6] by Gray and Stockham.

In Sec. 8 we consider the effect of sampling on the spectra of wobbled and non-wobbled sinusoids. Here we restrict to the case that the number of samples N is an integer multiple of the number M of periods of the sinusoid. Again, as in Secs. 5 and 6, we consider as a reference case the perfect quantizer ($h(x) = x$, no noise). For the case that no wobbling is applied, it turns out that one has to distinguish between three regimes, viz.

$$(4.8)\qquad 2\pi MA_0 < \frac{1}{C}\,N \; ,$$

$$(4.9)\qquad 2\pi MA_0 > CN \; ,$$

$$(4.10)\qquad 2\pi MA_0 \approx N \; ,$$

respectively, where C is a number somewhat larger than 1. These three cases correspond to what one may call the relatively oversampled, undersampled and critically sampled case, respectively. When (4.8) holds one has that the THD^2 behaves pretty much the same as a function of M and A_0, c_0 as it does for the time-continuous case. Explicitly, one has in this case for the maximal and average THD^2 (recall the meaning of L from (2.3))

$$(4.11)\qquad \mathrm{THD}^2_{\max} \approx \frac{16(L-1)}{\pi^2 A_0^3}\,\zeta^2_{-\frac{1}{2}}(0) = \frac{0.07006(L-1)}{A_0^3} \; ,$$

$$\overline{\mathrm{THD}}^2 \approx \frac{L-1}{\pi^4 A_0^3}\,\zeta_3(1) = \frac{0.01234(L-1)}{A_0^3} \tag{4.12}$$

with $\zeta_{-\frac{1}{2}}(0)$ as in (2.5) (also see (5.6)) and $\zeta_3(1) = \sum_1^\infty m^{-3}$. When (4.9) holds there is dominant folding-back of high-frequency components to the low-frequency range that is used in the definition of the THD^2. Thus it turns out that the THD^2 is rather insensitive to variations in A_0, c_0, and its average value is given by

$$\overline{\mathrm{THD}}^2 \approx \frac{(L-1)\,M}{3NA_0^2}\ , \tag{4.13}$$

as long as the condition $2LM < N$ is satisfied. In the transition region (4.10) the THD^2 behaves irregularly as a function of A_0, c_0, with values that are significantly larger than those obtained when the behaviour and magnitude of the THD^2 from either range (4.8) or (4.9) were extrapolated to the range (4.10). It is, furthermore, indicated that for the SINAD^{-2} one has to distinguish between the regimes (4.8–10) in a similar fashion.

For the case that wobbling with the ramp is done, the various ranges and respective behaviour and magnitude of the THD^2 can be obtained from those of the non-wobbled case simply by replacing A_0 by MA_0 and c_0 by Mc_0. Hence, the observation made above that wobbling amounts qualitatively to adding ${}^2\log M$ bits to the quantizer remains in force for the case of sampled sinusoids. We thus see that, when the number of samples N and the amplitude A_0 are kept fixed, increasing M (while respecting the condition that N/M is integer) results in a decrease of the rounding THD^2 by a factor of M^{-3} and M^{-1} in the ranges of relative oversampling and undersampling, respectively. The take-over point, defined as the point M where

$$\frac{D(L-1)}{M^3 A_0^3} = \frac{L-1}{3NMA_0^2}\ , \tag{4.14}$$

with D a number between 0.01234 and 0.07006, see (4.11–12), is close to $M = (N/2\pi A_0)^{1/2}$. However, around this take-over point the THD^2 behaves erratic and is considerably larger than either side of (4.14).

In Sec. 9 we present details about the formula that reconstructs the non-linearity h in terms of the harmonics of f^{sw} in (4.5) with f of the form (4.4). We take into account that the latter harmonics, due to the wobbling, may not properly reflect the symmetries present in the signal $h(A_0 \sin 2\pi Mt - c_0)$.

In Sec. 10 we present the results of simulations. We shall thus verify the above made observations on the THD^2 due to rounding for a perfect quantizer using a non-wobbled or a subtractively wobbled sinusoid as input and with finite sample sizes. We furthermore generate for a B-bits quantizer a polynomial non-linearity h of degree $\leq L$ whose contribution to the THD^2 is of the same order or smaller than the THD^2 that would result from the rounding operation only. We also give a comparison with the results of reconstruction when (subtractive or non-subtractive) dithering is used for stabilizing the harmonics from which the estimate of h is built.

We shall often be short about the proofs of our results. As already said, all results, and many more with more details, occur in the publicly available report [7].

5. The Fourier spectrum of a quantized sinusoid

In this section we give analytic results for the Fourier coefficients b_n of the quantization error

$$e(t) = [A_0 \sin 2\pi M t - c_0] - (A_0 \sin 2\pi M t - c_0) = \sum_{n=-\infty}^{\infty} b_n \, e^{2\pi i n M t} \,. \tag{5.1}$$

Here M is a positive integer and the amplitude A_0 is normally large compared to the offset c_0. We pay particular attention to the asymptotics of the b_n when $A_0 \to \infty$ and n, c_0 are bounded, and we present consequences for the THD^2. We give, furthermore, a somewhat heuristic and qualitative analysis on the global behaviour of the spectrum. We determine the asymptotic form of the average quantization error power, and we present generalizations of some of the results for quantization errors involving more genereal M^{-1}-periodic signals, such as the f in (4.4) with h a mild non-linearity. Finally, we indicate that for signals f of the latter type, the THD^2 is additive with respect to the distortion due to the non-linearity h and the rounding distortion.

5.1. Bessel series representation, asymptotics as $A_0 \to \infty$ and consequences for the THD^2

We have

$$b_n = \sum_{m \neq 0} \frac{(-1)^m \, J_n(2\pi m A_0)}{2\pi i m} \, e^{-2\pi i m c_0} \,, \qquad n \in \mathbb{Z} \,, \tag{5.2}$$

where J_n is the n^{th} Bessel function of the first kind and integer order n, see [1], Ch. 9. We refer to [7], Subsec. 2.1, for a proof and a literature survey concerning the result (5.2) and for expressions of the b_n's in terms of inverse trigonometric functions and Chebyshev polynomials.

By using the asymptotics of the Bessel functions for finite order and large argument, see [1], Ch. 9, and $J_n(-z) = (-1)^n \, J_n(z)$, we get the approximation for fixed $n \in \mathbb{Z}$, $c_0 \in \mathbb{R}$ as $A_0 \to \infty$

$$b_n \approx \frac{\sqrt{2}}{\pi \sqrt{A_0}} \{(-i)^n \, \zeta_{-\frac{1}{2}}(A_0 - c_0 - \tfrac{1}{2}) - i^n \, \zeta_{-\frac{1}{2}}(A_0 + c_0 - \tfrac{1}{2})\} \,. \tag{5.3}$$

Here $\zeta_{-\frac{1}{2}}(x)$ is the Hurwitz zeta function, see [11], Ch. 13,

$$\zeta_{-\frac{1}{2}}(x) = \frac{1}{2\pi \sqrt{2}} \sum_{m=1}^{\infty} \frac{\sin(2\pi m x - \frac{1}{4}\pi)}{m^{\frac{3}{2}}} \,, \qquad x \in \mathbb{R} \,, \tag{5.4}$$

a plot of which is presented in Fig. 1. We refer to [7], Subsec. 2.2 for a proof and a literature survey concerning the result (5.3–4), a full asymptotic series expansion of the b_n's involving higher order Hurwitz zeta function $\zeta_{k-\frac{1}{2}}$, $k \in \mathbb{Z}$, $k \leq 0$, and an efficient method for computation of these zeta functions.

As a consequence of (5.3–4) we have for the THD^2 based upon the frequencies nM, $n = 2, ..., L$ (where we use that $b_{-n} = b_n^*$, $n \in \mathbb{Z}$),

$$\text{(5.5)} \qquad \mathrm{THD}^2 = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |b_n|^2 \ ,$$

in good approximation

$$\text{(5.6)} \qquad \mathrm{THD}^2 \approx \frac{8(L-1)}{\pi^2 A_0^3} \{\zeta^2_{-\frac{1}{2}}(A_0 - c_0 - \tfrac{1}{2}) + \zeta^2_{-\frac{1}{2}}(A_0 + c_0 - \tfrac{1}{2})\} \ .$$

In particular, we find for the maximum value $\mathrm{THD}^2_{\max}$ (maximum over A_0, c_0 and using that $|\zeta_{-\frac{1}{2}}(x)|$ is maximal at $x = 0$)

$$\text{(5.7)} \qquad \mathrm{THD}^2_{\max} \approx \frac{16(L-1)}{\pi^2 A_0^3}\, \zeta^2_{-\frac{1}{2}}(0) = \frac{0.07006(L-1)}{A_0^3} \ ,$$

and for the average value $\overline{\mathrm{THD}^2}$ (average over a unit interval of A_0 and c_0) we get

$$\text{(5.8)} \qquad \overline{\mathrm{THD}^2} \approx \frac{L-1}{\pi^4 A_0^3} \sum_{m=1}^{\infty} \frac{1}{m^3} = \frac{0.01234(L-1)}{A_0^3} \ .$$

We refer to [7], Subsec. 2.3, for more details concerning (5.7–8). Observe that $\mathrm{THD}^2_{\max}$ and $\overline{\mathrm{THD}^2}$ differ by a factor of 6 (that is, 7.5 dB), and this accounts for the large fluctuations in the THD^2-values when A_0, c_0 are varied only slightly.

5.2. Global behaviour of the spectrum

We now present some observations on the global behaviour of the Fourier spectrum of $e(t)$ for the purpose of studying the effect of sampling as is done in Sec. 8. We have now to consider the $|b_n|^2$ also for large n so that the asymptotics in (5.3) does not apply. To that end we define

$$\text{(5.9)} \qquad S(n) := \int_{-\frac{1}{2}}^{\frac{1}{2}} |b_n(A_0, c_0)|^2 \, dc_0 \ , \qquad n \geq 0 \ ,$$

where we have explicitly indicated the dependence of b_n on A_0 and c_0. By (5.2) there holds

$$(5.10)\qquad S(n)=\sum_{m=1}^{\infty} S_m(n)\ ;\qquad S_m(n)=\frac{J_n^2(2\pi m A_0)}{2\pi^2 m^2}\ .$$

Using the asymptotics of the Bessel functions for large order and large arguments, see [1], Subsec. 9.3, it turns out that the effective value of S_m, loosely defined as

$$(5.11)\qquad \overline{S_m(n)}:=\frac{1}{2\Delta n}\int_{n-\Delta n}^{n+\Delta n} S_m(x)\,dx\ ,$$

is given in good approximation by

$$(5.12)\qquad \overline{S_m(n)}\approx\begin{cases}\dfrac{1}{2\pi^3 m^2}\min\left\{\dfrac{1}{\sqrt{(2\pi m A_0)^2-n^2}}\,,\,\dfrac{c\pi}{(\pi m A_0)^{\frac{2}{3}}}\right\}\ , & n\le 2\pi m A_0\ ,\\ 0\ , & n>2\pi m A_0\ ,\end{cases}$$

where

$$(5.13)\qquad c=\max_x\ Ai^2(x)=0.2870$$

with Ai the Airy function, see [1], Subsec. 10.4. For more details, we refer to [7], Subsec. 2.4 (we have carried through some corrections here). By illustration we have plotted in Fig. 2 for $m=3$ and $A_0=32$ the quantity $2\pi^2m^2S_m(n)$ (dots at the integer values of the argument n) together with $2\pi^2m^2\,\overline{S_m(n)}$ as approximated using (5.12) (solid line for the real values of the argument n). Evidently, $\overline{S_m(n)}$ gives very well the global picture of the behaviour of $S_m(n)$.

It follows that, away from the points $n=2\pi m A_0$, $m=1,2,\ldots$, the series in (5.10) for $S(n)$ is effectively over those m that satisfy $2\pi m A_0>n$, with the effective value given by

$$(5.14)\qquad \overline{S(n)}\approx\frac{1}{4\pi^4A_0}\sum_{2\pi m A_0>|n|}\frac{1}{m^3}\,\frac{1}{\sqrt{1-\left(\dfrac{n}{2\pi m A_0}\right)^2}}\ .$$

This agrees with the findings in [5], except that the authors of [5] declare (5.14) to be valid for all n, with a somewhat doubtful argumentation as to why $\overline{S(n)}$ should remain finite at the points $n=2\pi m A_0$, $m=1,2,\ldots$. As an amusing fact we note that when the right-hand side of (5.14) is summed over all $n\in\mathbb{Z}$ and $A_0\to\infty$, we obtain $\frac{1}{12}$, the well-known average value of the quantization error power with unit quantization step.

In Fig. 3 we have plotted

$$(5.15)\qquad \overline{S(n)}=\sum_{m=1}^{\infty}\overline{S_m(n)}$$

as a function of $n\ge 0$ on a log-scale ($\overline{S_m(n)}$ as approximated by the right-hand side of (5.12)) for $A_0=32$. We thus see a sort of staircase, with relatively flat regions between the points

$n = 2\pi m A_0$, $m = 1, 2, \ldots$, slowly rising before and sharply dropping after these points. We refer to [7], Subsec. 2.4, for more details concerning this staircase.

Obviously, when $n < 2\pi A_0$, the first term in (5.15), see the right-hand side of (5.14), dominates the others by far. It also follows from (5.14) that for small n we have

$$\overline{S(n)} \approx \frac{1}{4\pi^4 A_0} \sum_{m=1}^{\infty} \frac{1}{m^3} \, , \tag{5.16}$$

which is in complete agreement with what one finds for $\overline{\mathrm{THD}^2}$ in (5.8).

5.3. The SINAD and average error power

The SINAD for a perfect quantizer with input $A_0 \sin 2\pi M t - c_0$ is given by

$$\mathrm{SINAD}^{-2} = \frac{1}{\frac{1}{2}A_0^2} \int_0^1 e^2(t)\, dt = \frac{1}{\frac{1}{2}A_0^2} \sum_{n=-\infty}^{\infty} |b_n|^2 \, . \tag{5.17}$$

In [7], Subsec. 2.5, it is shown that

$$\int_0^1 e^2(t)\, dt = \frac{1}{12} + \sum_{m=1}^{\infty} \frac{(-1)^m \, J_0(2\pi m A_0)}{\pi^2 m^2} \cos 2\pi m c_0 \approx \tag{5.18}$$

$$\approx \frac{1}{12} - \frac{4\sqrt{2}}{3\pi\sqrt{A_0}} \{\zeta_{-\frac{3}{2}}(A_0 - c_0 - \tfrac{1}{2}) + \zeta_{-\frac{3}{2}}(A_0 + c_0 - \tfrac{1}{2})\} \, ,$$

the last approximation holding as $A_0 \to \infty$. Here $\zeta_{-\frac{3}{2}}$ is another Hurwitz zeta function, given by

$$\zeta_{-\frac{3}{2}}(x) = \frac{-3}{8\pi^2\sqrt{2}} \sum_{m=1}^{\infty} \frac{\cos(2\pi m x - \frac{1}{4}\pi)}{m^{\frac{5}{2}}} \, , \qquad x \in \mathbb{R} \, , \tag{5.19}$$

a plot of which is presented in Fig. 1.

We refer to [7], Subsec. 2.4, for expressions for all Fourier coefficients of $e^2(t)$ (not just the DC-term) in terms of inverse trigonometric functions and Chebyshev polynomials, in terms of Bessel functions, their asymptotics as $A_0 \to \infty$ in terms of Hurwitz zeta functions, an efficient way for computing $\zeta_{-\frac{3}{2}}$, as well as a literature survey on results and expressions as the ones given above.

5.4. More general signals and additivity of THD^2

We consider now more general quantization error signals

(5.20) $$e(t) = [f(t)] - f(t) = \sum_{n=-\infty}^{\infty} b_n\, e^{2\pi i n M t}$$

with $f(t)$ of the form

(5.21) $$f(t) = A \sin 2\pi M t - c(t)\ ,$$

where $c(t)$ is a smooth, M^{-1}-periodic signal of relatively small amplitude compared to A. In [7], Subsecs. 5.2, 5.5, it is shown that the asymptotic form of the b_n's and the average power of $e(t)$ as $A \to \infty$ have the same mathematical form as the right-hand sides of (5.3) and (5.18), except that the arguments $A_0 \mp c_0 - \frac{1}{2}$ in the latter expressions should be replaced by the arguments $A \mp c\Big(\pm \dfrac{1}{4M}\Big) - \frac{1}{2}$. Observe that these latter arguments are, apart from the $-\frac{1}{2}$, exactly the extreme values of $f(t)$.

For $f(t)$ of the form

(5.22) $$f(t) = h(A_0 \sin 2\pi M t - c_0) = \sum_{n=-\infty}^{\infty} c_n\, e^{2\pi i n M t}\ ,$$

where h is a mild non-linearity, we define the total THD^2 by

(5.23) $$\mathrm{THD}^2_{\mathrm{tot}} = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |d_n|^2\ ; \qquad [f(t)] = \sum_{n=-\infty}^{\infty} d_n\, e^{2\pi i n M t}\ .$$

It is shown in [7], end of Subsec. 2.6, that this $\mathrm{THD}^2_{\mathrm{tot}}$ is in good approximation additive with respect to the THD^2 due to rounding and the THD^2 due to h. More precisely, we have

(5.24) $$\mathrm{THD}^2_{\mathrm{tot}} \approx \mathrm{THD}^2_{\mathrm{rounding}} + \mathrm{THD}^2_h\ ,$$

where

(5.25) $$\mathrm{THD}^2_h = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |c_n|^2\ ,$$

and $\mathrm{THD}^2_{\mathrm{rounding}}$ is the rounding THD^2 of Subsec. 5.1 for a sinusoidal input signal $A' \sin 2\pi M t - c'$ with A' of the same order of magnitude as A and c' small compared to A'.

6. Wobbling

In this section we consider M^{-1}-periodic signals

(6.1) $$f(t) = h(A_0 \sin 2\pi M t - c_0)$$

with h a mild non-linearity as before, and we want to find reliable estimates of the Fourier coefficients of $f(t)$ from a rounded version of $f(t)$ that are relatively insensitive to variations in amplitude A_0 and offset c_0. We propose to achieve this by constructing an additive signal $g(t)$, $0 \leq t < 1$, such that the harmonics of the error signal

$$e^{sw}(t) = [f(t) + g(t)] - (f(t) + g(t)) = ([f(t) + g(t)] - g(t)) - f(t) \tag{6.2}$$

corresponding to $f(t) + g(t)$ have decreased in power as much as possible. Because of the far right-hand expression in (6.2) for $e^{sw}(t)$, one can state this problem also as choosing $g(t)$ such that the harmonics of the subtractively wobbled signal

$$f^{sw}(t) = [f(t) + g(t)] - g(t) \tag{6.3}$$

are as close to those of f itself as possible.

In [7], Sec. 3, it is shown that the g's with best average performance in achieving this goal satisfy the following uniform distribution property: for any $t \in [0, M^{-1})$ there is an $\alpha(t) \in [0, M^{-1})$ such that

$$\Big\{ g\Big(t + \frac{p}{M}\Big) (\mathrm{mod}\, \mathbb{Z}) \mid p = 0, ..., M-1 \Big\} = \Big\{ \alpha(t) + \frac{p}{M} \mid p = 0, ..., M-1 \Big\} . \tag{6.4}$$

Among the many functions $g(t)$, $0 \leq t < 1$, that satisfy this property are the ramps $r_Q(t)$ with integer span Q such that $\gcd(Q, M) = 1$:

$$g(t) = r_Q(t) = Q(t - \tfrac{1}{2}) \;, \qquad 0 \leq t < 1 \;. \tag{6.5}$$

The effect of wobbling using such a ramp can be read off from the formula

$$\frac{1}{M} \sum_{p=0}^{M-1} \Big(\Big[x + Q\,\frac{p}{M} - \frac{1}{2} \Big] - \Big(x + Q\,\frac{p}{M} - \frac{1}{2} \Big) \Big) = \frac{1}{M} \left([Mx - \tfrac{1}{2}] - (Mx - \tfrac{1}{2}) \right) , \tag{6.6}$$

see [7], (4.22). This formula (6.6) is basic to the observation that wobbling using r_Q amounts to reducing the stepsize of the quantizer by a factor of M.

In the remainder of this section we shall mainly consider wobbling with the standard ramp $r(t)$, $0 \leq t < 1$, for which the choice $Q = 1$ is made in (6.5), and we shall examine what form the results of Sec. 5 take when wobbling is used. Hence, we consider the input sinusoid

$$f(t) = A_0 \sin 2\pi M t - c_0 \tag{6.7}$$

and the associated error signal after subtractive wobbling

$$e^{sw}(t) = [f(t) + r(t)] - (f(t) + r(t)) = \sum_{k=-\infty}^{\infty} b_k^{sw}\, e^{2\pi i k t} \;. \tag{6.8}$$

Note that wobbling causes the error signal $e^{sw}(t)$ to have non-zero frequency components at all frequencies $k \in \mathbb{Z}$ and not just at the harmonic frequencies $k = nM, n = \mathbb{Z}$. We denote by

$$(6.9) \qquad H_n^{sw} = b_{nM}^{sw} \ , \qquad n \in \mathbb{Z} \ ,$$

the Fourier coefficients of $e^{sw}(t)$ at the harmonic frequencies.

6.1. Bessel series representation, asymptotics as $A_0 \to \infty$ and consequences for the THD^2

We have for $n, m \in \mathbb{Z}$

$$(6.10) \qquad b_{nM+m}^{sw} = \sum_{l=-\infty\,;\,lM+m\neq 0}^{\infty} \frac{J_{n-l}(2\pi(lM+m)\,A_0)}{2\pi i(lM+m)}\, e^{-2\pi i(lM+m)c_0}$$

with J_{n-l} Bessel functions as before. See [7], Subsec. 4.2.

By using the asymptotics of the Bessel functions, see [7], Subsec. 4.1, for the details, we have for fixed functions $n \in \mathbb{Z}, c_0 \in \mathbb{R}$ as $A_0 \to \infty$

$$(6.11) \qquad H_n^{sw} = b_{nM}^{sw} = \sum_{l\neq 0} \frac{J_{n-l}(2\pi lMA_0)}{2\pi ilM}\, e^{-2\pi ilMc_0} \approx$$

$$\approx \frac{1}{M^{\frac{3}{2}}}\, \frac{\sqrt{2}}{\pi\sqrt{A_0}}\, \{(-i)^n\, \zeta_{-\frac{1}{2}}(M(A_0 - c_0) + \tfrac{1}{4}) - i^n\, \zeta_{-\frac{1}{2}}(M(A_0 + c_0) + \tfrac{1}{4})\}$$

with $\zeta_{-\frac{1}{2}}$ the Hurwitz zeta function of (5.4). Note that the far right-hand side of (6.11) has the same mathematical form as the right-hand side of (5.3), except for the factor $M^{-\frac{3}{2}}$ and the M in the arguments of $\zeta_{-\frac{1}{2}}$.

As a consequence of (6.11), we have for the THD^2 based upon frequencies nM, $n = 2, ..., L$ (using $H_{-n}^{sw} = (H_n^{sw})^*$),

$$(6.12) \qquad \mathrm{THD}^{2,sw} = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |H_n^{sw}|^2 \ ,$$

in good approximation

$$(6.13) \qquad \mathrm{THD}^{2,sw} \approx \frac{8(L-1)}{\pi^2 (MA_0)^3}\, \{\zeta^2_{-\frac{1}{2}}(M(A_0 - c_0) + \tfrac{1}{4}) + \zeta^2_{-\frac{1}{2}}(M(A_0 + c_0) + \tfrac{1}{4})\} \ .$$

The maximum and average value of THD^{2,sw} are given in good approximation by

$$(6.14) \qquad \mathrm{THD}^{2,sw}_{\max} \approx \frac{16(L-1)}{\pi^2 (MA_0)^3}\, \zeta^2_{-\frac{1}{2}}(0) = \frac{0.7006(L-1)}{(MA_0)^3} \ ,$$

$$\text{(6.15)} \qquad \overline{\mathrm{THD}^{2,sw}} \approx \frac{L-1}{\pi^4 (MA_0)^3} \sum_{m=1}^{\infty} m^{-3} = \frac{0.01234(L-1)}{(MA_0)^3} \; .$$

The formulas (6.13–15), see (5.6–8), show that wobbling with a ramp covering M sine periods amounts to replacing the amplitude A_0 and offset c_0 by MA_0 and Mc_0, respectively.

6.2. Global behaviour of the harmonics of quantized wobbled sinusoids

The observations made in Subsec. 5.2 on the global behaviour of the spectrum of $e(t)$ continue to hold for the harmonic part of the spectrum of $e^{sw}(t)$ with little changes, except that the amplitude A_0 should be replaced by MA_0. Indeed, letting

$$\text{(6.16)} \qquad S^{sw}(n_) := \int_{-\frac{1}{2}}^{\frac{1}{2}} |H_n^{sw}(A_0, c_0)|^2 \, dc_0 \; , \qquad n \geq 0 \; ,$$

we have from the first identity in (6.11) that

$$\text{(6.17)} \qquad S^{sw}(n) = \sum_{m=1}^{\infty} \frac{J_{n-m}^2(2\pi m M A_0) + J_{n+m}^2(2\pi m M A_0)}{4\pi^2 m^2 M^2} \; .$$

Since only small m contribute substantially to the right-hand side series in (6.17) while the effective values of $J_{n\pm m}^2(z)$ and $J_n^2(z)$ are largely the same when n is large and m is small, it follows that $S^{sw}(n)$ has an effective value which is the same as that of $S(n)$ in (5.10), with A_0 replaced by MA_0.

6.3. The SINAD, average error power and redistribution of error power

The SINAD^{sw} for a perfect quantizer with input $A_0 \sin 2\pi Mt - c_0$ is given by

$$\text{(6.18)} \qquad \mathrm{SINAD}^{-2,sw} = \frac{1}{\frac{1}{2}A_0^2} \int_0^1 (e^{sw}(t))^2 \, dt = \frac{1}{\frac{1}{2}A_0^2} \sum_{k=-\infty}^{\infty} |b_k^{sw}|^2 \; .$$

In [7], Subsecs. 4.2, 4.6, it is shown that (carrying through a minor correction) as $A_0 \to \infty$

$$\text{(6.19)} \qquad \int_0^1 (e^{sw}(t))^2 \, dt =$$

$$= \frac{1}{12} + \sum_{m \neq 0} \frac{J_{-m}(2\pi m M A_0)}{2\pi^2 m^2 M^2} \, e^{-2\pi i m M c_0} \approx$$

$$\approx \frac{1}{12} - \frac{1}{M^{\frac{5}{2}}} \frac{4\sqrt{2}}{3\pi\sqrt{A_0}} \{ \zeta_{-\frac{3}{2}}(M(A_0 - c_0) + \tfrac{1}{4}) + \zeta_{-\frac{3}{2}}(M(A_0 + c_0) + \tfrac{1}{4}) \} \; ,$$

with $\zeta_{-\frac{3}{2}}$ the Hurwitz zeta function of (5.19). Note that the second term in the far right-hand side of (6.19) has the same mathematical form as the second term in the far right-hand side of (5.18), except for the factor $M^{-\frac{5}{2}}$ and the M in the arguments of $\zeta_{-\frac{3}{2}}$.

In [7], Subsec. 4.2, it is shown that for any M^{-1}-periodic signal f (and, whence, not only for f in (6.7)) there holds

$$(6.20) \qquad \sum_{n=-\infty}^{\infty} |b_{nM+m}^{sw}|^2 = \frac{1}{4M^2 \sin^2 \dfrac{\pi m}{M}} \, , \qquad m = 1, ..., M-1 \, .$$

Note that the right-hand side of (6.20) is independent of f. Moreover, there holds

$$(6.21) \qquad \sum_{n=-\infty}^{\infty} |H_n^{sw}|^2 - \frac{1}{12M^2} = \sum_{k=-\infty}^{\infty} |b_k^{sw}|^2 - \frac{1}{12} \, ,$$

and either side of (6.21) is of order $M^{-\frac{5}{2}}$ on account of (6.18) and (6.19). Hence, also see (5.17–18), the total error power at the harmonics has been reduced by a factor $1/M^2$ due to wobbling with a ramp covering M sine periods. The error power at the harmonics present in the non-wobbled error signal $e(t)$ has been largely removed to the interharmonic frequencies $nM + m, n \in \mathbb{Z}, m = 1, ..., M-1$, where we also note that the right-hand sides of (6.20) sum to $\frac{1}{12}(1 - M^{-2})$.

6.4. Some generalizations

Many of the previous results of this section continue to hold when the unit span ramp $r(t)$ is replaced by the ramp $r_Q(t)$ of span Q with $\gcd(Q, M) = 1$, see (6.5). This can be based on the Bessel series representation

$$(6.22) \qquad H_n^{sw} = \sum_{m \neq 0} \frac{(-1)^{mM(Q-1)} \, J_{n-mQ}(2\pi m M A_0)}{2\pi i m M} \, e^{-2\pi i m M c_0}$$

for the corresponding harmonic components, see [7], Subsec. 4.5, for more details.

A further generalization consists of allowing more general signals

$$(6.23) \qquad f(t) = h(A_0 \sin 2\pi M t - c_0) = A \sin 2\pi M t - c(t)$$

as was done in Subsec. 5.4. Indeed, there do exist asymptotic formulas of the type (6.11) and (6.19) in which the $A_0 \mp c_0$ occurring in the arguments of the Hurwitz zeta functions should be replaced by the maximum and minimum value $f\left(\pm \dfrac{1}{4M}\right)$ of f, see [7], Subsec. 4.6. Furthermore, an approximate additivity of $\mathrm{THD}_{\mathrm{tot}}^{2,sw}$ with respect to $\mathrm{THD}_{\mathrm{rounding}}^{2,sw}$ and THD_h^2, see Subsec. 5.4, holds in this more general situation.

7. The SINAD for noisy quantizers; relation with the theory of non-subtractive noise dither

We now mention some results for the SINAD in case of a noisy quantizer, and we establish a connection with the theory of non-subtractive noise dither. As usual we consider M^{-1}-periodic signals f of the form

$$f(t) = h(A_0 \sin 2\pi Mt - c_0) = A \sin 2\pi Mt - c(t) \tag{7.1}$$

where

$$c(t) = \sum_{n=-\infty}^{\infty} c_n \, e^{2\pi i n M t} \tag{7.2}$$

is a smooth, M^{-1}-periodic signal of relatively small amplitude compared to A. Because of the special form of f we can assume that $c_{\pm 1} = 0$.

We consider a noisy quantizer in (1.1) and we distinguish between the cases that wobbling is or is not applied according to

$$f^{sw}(t) = [f(t) + \mathbf{n}(t) + r(t)] - r(t) \ , \qquad \text{or } f^{nw}(t) = [f(t) + \mathbf{n}(t)] \ . \tag{7.3}$$

With $a = sw, nw$ we define the SINAD^a by

$$\text{SINAD}^{-2,a} = \frac{1}{\frac{1}{2}(A^a)^2} \int_0^1 |f^a(t) - A^a \sin 2\pi Mt|^2 \, dt \ , \tag{7.4}$$

where

$$A^a = 2 \int_0^1 f^a(t) \sin 2\pi Mt \, dt \ . \tag{7.5}$$

Under the assumption that $\mathbf{n}(t)$ has zero mean and sufficiently small correlation times, it is shown in [7], Sec. 5, that

$$\text{SINAD}^{-2,a} = \frac{1}{\frac{1}{2}(A^a)^2} \left(E^a + \sum_n |c_n|^2 + \varepsilon^a \right) , \tag{7.6}$$

where E^a is the average errow power

$$E^a = \int_0^1 |f^a(t) - f(t)|^2 \, dt \tag{7.7}$$

and ε^a is a term involving the Fourier coefficients of the error signal appropriately averaged with the pdf of $\mathbf{n}$ as weighting function. Note that $\sum_n |c_n|^2$ is the same for either case $a = sw, nw$ and that it is closely related to the THD^2 due to the non-linearity h. Indeed, when h is a polynomial of degree $\leq L$ and $c_0 = 0$ we have, see (5.2), that these two quantities coincide, except for a factor $\frac{1}{4} A_0^2$.

It is argued in [7], Sec. 5, that ε^a can be ignored, especially when $a = sw$. Furthermore, for E^a there is shown that

$$(7.8) \qquad E^a = \frac{1}{12} + \sigma^2 + \sum_{l \neq 0} P_f^a(-l)\, K_l^{(2)} .$$

Here σ^2 is the variance of the noise, $K_l^{(2)}$ is obtained by setting $k = 2$ in

$$(7.9) \qquad K_l^{(k)} = \int_0^1 E_{\mathbf{n}}(|[y + \mathbf{n}] - y|^k)\, e^{-2\pi i l y}\, dy = \\ = \Big(\frac{-1}{2\pi i}\Big)^k \Big(\frac{\sin \pi\nu}{\pi\nu}\, P_{\mathbf{n}}(\nu)\Big)^{(k)} (\nu = l)$$

with $P_{\mathbf{n}}(\nu) = \int e^{-2\pi i \nu x}\, p_{\mathbf{n}}(x)\, dx$ the characteristic function of the noise, and

$$(7.10) \qquad P_f^{sw}(\nu) = \frac{\sin \pi\nu}{M \sin \dfrac{\pi\nu}{M}} \int_0^1 e^{-2\pi i \nu f(t)}\, e^{-2\pi i \nu (t - \frac{1}{2})/M}\, dt ,$$

$$(7.11) \qquad P_f^{nw}(\nu) = \int_0^1 e^{-2\pi i \nu f(t)}\, dt .$$

Hence, the quantities $P_f^a(-l)$ in (7.8) are for both cases $a = sw, nw$ of similar nature, except that $P_f^{sw}(-l)$ vanishes unless l is a multiple of M whereas $P_f^{nw}(-l) \neq 0$ in general for all l.

It follows that, as M gets large, we can ignore both the ε^{sw} at the right-hand side of (7.6) and the right-hand side series in (7.8) and we obtain

$$(7.12) \qquad \mathrm{SINAD}^{-2,sw} = \frac{1}{\frac{1}{2}(A^{sw})^2} \Big(\frac{1}{12} + \sigma^2 + \sum_n |c_n|^2\Big) .$$

This then shows that the rounding distortion ($\frac{1}{12}$), the variance of the noise (σ^2) and the THD^2 (represented by $\sum_n |c_n|^2$) occur additively in the right-hand side of (7.12) which is intuitively very pleasing.

It is, furthermore, interesting to reverse the perspective with respect to the noise as follows. Assume that we have a noiseless quantizer and that we deliberately add dither noise $\mathbf{n}_d$ to the input signal. Hence in the above we replace $\mathbf{n}$ by $\mathbf{n}_d$. By choosing the pdf of the dither noise such that $K_l^{(2)}$ vanishes for all $l \neq 0$ (the far right-hand side of (7.9) tells us how this can be achieved), we see that the right-hand side series in (7.8) vanishes altogether so that E^a becomes independent of f. The simplest dither noise $\mathbf{n}_d$ for which $K_l^{(2)}$ vanishes for $l \neq 0$ has a triangular pdf on $[-1,1]$ for which $\sigma^2 = 1/6$. This (and, more generally, formula (7.9)) establishes a relation with some basic results in [6] from the theory of non-subtractive noise dither. Note that the results in this section deal with the case of time-continuous input and output signals and that nothing has been said about how and to what extent sampling has a deteriorating influence.

8. Wobbling and sampling

In this section we present the DFT-analysis of quantized wobbled and non-wobbled sinusoids and signals of the form

$$(8.1) \qquad f(t) = h(A_0 \sin 2\pi Mt - c_0) = A \sin 2\pi Mt - c(t)$$

with h and c as before, We let

$$(8.2) \qquad f^{sw}(t) = [f(t) + r(t)] - r(t) \; , \qquad f^{nw}(t) = [f(t)] \; ,$$

with $r(t)$ the standard ramp, and error signals

$$(8.3) \qquad e^{sw}(t) = f^{sw}(t) - f(t) \; , \qquad e^{nw}(t) = f^{nw}(t) - f(t) \; .$$

With $a = sw, nw$ we have

$$(8.4) \qquad \int\limits_0^1 f^a(t)\, e^{-2\pi inMt}\, dt = \int\limits_0^1 f(t)\, e^{-2\pi inMt}\, dt + \varepsilon_n^a \; ,$$

where

$$(8.5) \qquad \varepsilon_n^a = \int\limits_0^1 e^a(t)\, e^{-2\pi inMt}\, dt$$

are the H_n^{sw}'s of Sec. 6 and the b_n's of Sec. 5, respectively.

We now let $N = KM$ with integer K, and we consider the DFT$_N$-versions of (8.4–5), where we use the definition

$$(8.6) \qquad \mathrm{DFT}_N[(y(k))_{k=0,\dots,N-1}]\,(l) = \frac{1}{N} \sum_{k=0}^{N-1} y(k)\, e^{-2\pi ikl/N} \; , \qquad l = 0, \dots, N-1 \; ,$$

for the DFT$_N$ of the sequence $(y(k))_{k=0,\dots,N-1}$. We have for $n = 0, \dots, K-1$, see [7], Sec. 6, that

$$\text{(8.7)} \qquad \mathrm{DFT}_N\Big[\Big(f\Big(\frac{k}{N}\Big)\Big)_{k=0,...,N-1}\Big]\,(l) = \mathrm{DFT}_K\Big[\Big(f\Big(\frac{k/K}{M}\Big)\Big)_{k=0,...,K-1}\Big]\,(n)\ .$$

When f is a trigonometric polynomial of degreee $L < \frac{1}{2}K$, we have that (8.7) equals the n^{th} harmonic of f as given by the first quantity at the right-hand side of (8.4). And in [7], Sec. 6, it is shown that for such an f we have for $n = 0, ..., K-1$

$$\text{(8.8)} \qquad \mathrm{DFT}_N\Big[\Big(f^a\Big(\frac{k}{N}\Big)\Big)_{k=0,...,N-1}\Big]\,(nM) = \int\limits_0^1 f(t)\, e^{-2\pi i n M t}\, dt + \varepsilon^a_{K,n}\ ,$$

where

$$\text{(8.9)} \qquad \varepsilon^a_{K,n} := \sum_{j=-\infty}^{\infty} \varepsilon^a_{n+jK}$$

with ε^a_{n+jK} given in (8.5).

We shall now consider the case of a perfect quantizer, so that $h(x) = x$ in (8.1), and we are interested in the THD^2 based upon the DFT_N-harmonics at frequencies $2M, ..., LM$. We first consider the case that $a = nw$, no wobbling, and we shall distinguish between the regions

$$\text{(8.10)} \qquad 2\pi M A_0 < \frac{1}{C}\, N\ , \qquad 2\pi M A_0 > CN\ ,$$

where C is a constant somewhat larger than 1.

In the first region in (8.10), the term with $j = 0$ in (8.8) equals $\varepsilon^{nw}_n = b_n$ and dominates the others by far, see Subsec. 5.2. Consequently, we then have $\varepsilon^{nw}_{K,n} \approx \varepsilon^{nw}_n = b_n$, and the resulting THD^2, together with the maximum value and its average value, are given by formulas (5.6–8). In particular, there is no dependence on N.

When we are in the second region in (8.10), the right-hand side of (8.9) is a series of several unrelated terms ε^{nw}_{n+jK} of comparable average power, and this holds for all $n = 0, ., ..., K-1$. It must then be expected that all $\varepsilon^{nw}_{K,n}$ have average powers of comparable size. Since, see [7], Sec. 6,

$$\text{(8.11)} \qquad \varepsilon^{nw}_{K,n} = \mathrm{DFN}_K\Big[\Big(e^{nw}\Big(\frac{k/K}{M}\Big)\Big)_{k=0,...,K-1}\Big]\,(n)\ , \qquad n = 0, ..., K-1\ ,$$

we have by Parseval's theorem that

$$\text{(8.12)} \qquad \sum_{n=0}^{K-1} |\varepsilon^{nw}_{K,n}|^2 = \frac{1}{K} \sum_{k=0}^{K-1} \Big|e^{nw}\Big(\frac{k}{N}\Big)\Big|^2\ .$$

The right-hand side of (8.12) is an average quantization error, and is, therefore, in good approximation equal to $\frac{1}{12}$. Since

$$\text{(8.13)} \qquad \mathrm{THD}^{2,nw} = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} |\varepsilon_{K,n}^{nw}| \ ,$$

is a series containing unrelatied quantities, the fluctuations of the individual $|\varepsilon_{K,n}^{nw}|^2$ are averaged out. Hence the THD^2 in (8.13) is in good approximation equal to its average value, and this average value is given by

$$\text{(8.14)} \qquad \overline{\mathrm{THD}^{2,nw}} = \frac{1}{\frac{1}{4}A_0^2} \cdot \frac{1}{12} \cdot \frac{L-1}{K} = \frac{(L-1)\,M}{3NA_0^2} \ .$$

We next consider the case $a = sw$, subtractive wobbling; now we have that $\varepsilon_n^{sw} = H_n^{sw}$. As noted in Subsec. 6.2 the global shape of the harmonic part of the spectrum of $e^{sw}(t)$ is the same as that of the non-wobbled error signal $e^{nw}(t)$, except that A_0 should be replaced by MA_0. Accordingly, we now distinguish between the regions

$$\text{(8.15)} \qquad 2\pi M^2 A_0 < \frac{1}{C}\,N \ , \qquad 2\pi M^2 A_0 > CN \ .$$

In the first region in (8.15) we have $\varepsilon_{K,n}^{sw} \approx \varepsilon_n^{sw} = H_n^{sw}$, and the THD^2, together with its maximum value and its average value, are given by formulas (6.13–15). In particular, there is no dependence on N. In case we are in the second region in (8.15), the THD^2 is in good approximation equal to its average value, the latter value being obtained from (8.14) by replacing A_0 by MA_0, see [7], Sec. 6, for the details. Hence, in this second region, we have

$$\text{(8.16)} \qquad \overline{\mathrm{THD}^{2,sw}} \approx \frac{L-1}{3NMA_0^2} \ .$$

There is a reasonably large gap between the two regions in (8.10) (and between the two regions in (8.15)). In this gap region neither of the formulas for the THD^2 holding in one of the regionsin (8.10) is valid to good approximation. When A_0 is equal or somewhat larger than $K/2\pi$, one gets severe folding back in (8.9) from the first peak region near frequencies $\pm 2\pi MA_0$ in the spectrum of $e^{nw}(t)$. And when A_0 is decreased below $K/2\pi$, one gets a sharp drop in the THD^2 since then all aliasing terms in the series (8.9) come from the region just beyond the first distinct peak in the spectrum of $e^{nw}(t)$.

We conclude this section with some comments. For general $h(x)$ in (8.1) it can be shown that the THD^2 is in good approximation additive with respect to the rounding distortion and the distortion due to the non-linearity h. This holds for both the wobbled and non-wobbled case and in either of the regions in (8.10) and (8.15).

The formulas (6.20–21) continue to hold for the sampled case, when altered appropriately. For instance, there holds for any f

$$\text{(8.17)} \qquad \sum_{n=0}^{K-1} \left| \mathrm{DFT}_N \left[\left(e^{sw}\left(\frac{k}{N}\right)\right)_{k=0,\ldots,N-1} \right] (nM+m) \right|^2 = \frac{1}{4M^2 \, \sin^2 \dfrac{\pi m}{M}}$$

for $m = 1, ..., M - 1$.

A DFT_N-analysis for the total error energies

$$(8.18) \qquad \frac{1}{N} \sum_{k=0}^{N-1} \Big| e^a\Big(\frac{k}{N}\Big)\Big|^2 = \mathrm{DFT}_N\Big[\Big(\Big|e^a\Big(\frac{k}{N}\Big)\Big|^2\Big)_{k=0,...,N-1}\Big]\,(0)$$

can be given in a quite similar way as in the above. As a result one finds again that one has to distinguish between the two cases in (8.10) and (8.15), the first cases being described by formulas (5.18) and (6.19), respectively. For the second cases in (8.10) and (8.15) one has that

$$(8.19) \qquad \mathrm{SINAD}^{-2,a} = \frac{1}{\frac{1}{2}A_0^2} \cdot \frac{1}{N} \sum_{k=0}^{N-1} \Big| e^a\Big(\frac{k}{N}\Big)\Big|^2 = \frac{1}{\frac{1}{2}A_0^2}\,(\tfrac{1}{12} + \Delta^a)\ ,$$

where Δ^a, considered as a function of A_0 and c_0, has zero mean and mean square root given in good approximation by

$$(8.20) \qquad \sqrt{\frac{M}{180N}}\ , \qquad \frac{1}{M^2}\sqrt{\frac{M}{180N}}\ ,$$

for the cases $a = nw, sw$, respectively.

9. The reconstruction formula for the non-linearity h

In this section we present a formula for estimating a mild non-linearity h from the harmonics

$$(9.1) \qquad F_n^a = \int_0^1 f^a(t)\, e^{-2\pi i n M t}\, dt\ , \qquad |n| \le L\ ,$$

with f^a as given in (8.1–2), and the DFT_N-versions, see (8.8),

$$(9.2) \qquad F_{K,n}^a = \mathrm{DFT}_N\Big[\Big(f^a\Big(\frac{k}{N}\Big)\Big)_{k=0,...,N-1}\Big]\,(nM)\ , \qquad |n| \le L\ .$$

Consider for simplicity the case that $c_0 = 0$ in (8.1) and that h is a polynomial of degree $\le L$. Then one can reconstruct $h(x)$ on the range $|x| \le A$ exactly from the numbers

$$(9.3) \qquad F_n = \int_0^1 h(A_0 \sin 2\pi M t)\, e^{-2\pi i n M t}\, dt\ , \qquad |n| \le L\ ,$$

according to the formula

$$(9.4) \qquad h(x) = F_0 + 2 \sum_{n=1}^{L} i^n \, F_n \, T_n(x/A_0) \; , \qquad |x| \le A_0 \; .$$

Here $T_n(y) = \cos(n \arccos y)$ is the n^{th} Chebyshev polynomial of the first kind, see [1], p. 775. For the proof, see [7], Sec. 7.

When h is not a polynomial of degreee $\le L$, the polynomial $g_0(x)$ at the right-hand side of (9.4) is best in the sense that, among all polynomials $g(x)$ of degree $\le L$, the quantity

$$(9.5) \qquad \frac{1}{\pi} \int_{-A_0}^{A_0} \frac{|h(x) - g(x)|^2}{\sqrt{A_0^2 - x^2}} \, dx$$

is minimal for $g = g_0$.

The Fourier coefficients F_n reflect the symmetries present in $h(A_0 \sin 2\pi M t)$ according to

$$(9.6) \qquad i^n \, F_n = i^{-n} \, F_{-n} \; , \qquad n \in \mathbb{Z} \; .$$

Due to wobbling (and, if present, noise) the estimates $F^a_{K,n}$ in (9.2) of F_n do not necessarily satisfy (9.6). To avoid complex-valued estimates of h, this must be forced by replacing $F^a_{K,n}$ by

$$(9.7) \qquad \tilde{F}^a_{K,n} = \mathrm{Re}\,[F^a_{K,n}] \; , \;\; n \text{ even} \; ; \qquad \tilde{F}^a_{K,n} = i \, \mathrm{Im}\,[F^a_{K,n}] \; , \;\; n \text{ odd} \; .$$

One thus arrives at the estimate

$$(9.8) \qquad \tilde{h}^a(x) = \tilde{F}^a_{K,0} + 2 \sum_{n=1}^{L} i^n \, \tilde{F}^a_{K,n} \, T_n(x/A_0) \; , \qquad |x| \le A_0 \; ,$$

for $h(x)$ on the range $|x| \le A_0$. The choice of these $\tilde{F}^a_{K,n}$ is best in the sense that, given the numbers $F^a_{K,n}$, the mean square distance

$$(9.9) \qquad \int_0^1 \Big| g(A_0 \sin 2\pi M t) - \sum_{n=-L}^{L} F^a_{K,n} \, e^{2\pi i n M t} \Big|^2 \, dt$$

is minimal among all real-valued functions g for $g = \tilde{h}^a$.

10. Simulations

In this section we present simulation results to confirm the observations on the THD^2 due to rounding with sinusoidal input before and after wobbling in dependence of the number of samples N for M sine periods. We also show simulation resuls for the retrieval of polynomial non-linearities h having a contribution to the total THD^2 of comparable or smaller size than the contribution to the total THD^2 of the rounding before wobbling. We compare the performance of the wobbling technique with that of (subtractive) dither methods, both for the issue of reducing rounding THD^2 and for the issue of retrieving a polynomial non-linearity.

We let B be the number of bits of the quantizer. Accordingly, we consider inputs $x \in \mathbb{R}$ with

$$(10.1) \qquad -\tfrac{1}{2}\,(2^B - 1) \leq x \leq \tfrac{1}{2}\,(2^B - 1)\ .$$

We let $L-1$, the number of higher harmonics from which the THD^2 is built, be equal to 9, and we consider N samples at sample points k/N, $k = 0, ..., N-1$ of input and output signals of period $1/M$ with N an integer multiple of M. The input sinusoid has amplitude

$$(10.2) \qquad A_0 = \tfrac{1}{2}\,A_{\mathrm{rel}}(2^B - 1)\ ; \qquad A_{\mathrm{rel}} \in [0.8, 1.0]\ .$$

In all cases the offset c_0 of the sinusoid is taken 0. When wobbling is applied we use the unit-span standard ramp $r(t)$. We consider 5^{th} degree polynomial non-linearities h given in the form

$$(10.3) \qquad h(x) = \tfrac{1}{2}\,(2^B - 1)(\alpha_0 x_{\mathrm{rel}}^5 + \alpha_1 x_{\mathrm{rel}}^4 + \alpha_2 x_{\mathrm{rel}}^3 + \alpha_3 x_{\mathrm{rel}}^2 + \alpha_4 x_{\mathrm{rel}} + \alpha_5)$$

with $x = \frac{1}{2}\,(2^B - 1)\,x_{\mathrm{rel}}$ and $|x_{\mathrm{rel}}| \leq 1$.

There are three relevant THD^2-numbers to be considered, viz. the THD^2 due to h, defined by

$$(10.4) \qquad \mathrm{THD}_h^2 = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} \Big|\mathrm{DFT}_N\Big[\Big(h\Big(A_0 \sin 2\pi k\,\frac{M}{N}\Big)\Big)_{k=0,\ldots,N-1}\Big]\,(nM)\Big|^2\ ,$$

and the THD^2 due to rounding before and after subtractive wobbling, defined by

$$(10.5) \qquad \mathrm{THD}^{2,a}_{\mathrm{rounding}} = \frac{1}{\frac{1}{4}A_0^2} \sum_{n=2}^{L} \Big|\mathrm{DFT}_N\Big[\Big(f^a\Big(\frac{k}{N}\Big)\Big)_{k=0,\ldots,N-1}\Big]\,(nM)\Big|^2$$

with $a = nw, sw$ and

$$(10.6) \qquad f^{nw}(t) = [f(t)]\ , \qquad f^{sw}(t) = [f(t) + r(t)] - r(t)$$

and $f(t) = A_0 \sin 2\pi M t$.

In Fig. 4 we have plotted as a function of A_0, see (10.2)

$$(10.7) \qquad 10 \log_{10} \mathrm{THD}^{2,nw}_{\mathrm{rounding}}$$

with

$$(10.8) \qquad N = 2048\ ; \qquad M = 2, 32\ ; \qquad B = 6\ .$$

In the case $M = 2$ we have that the first inequality in (8.10) is valid. Hence, as we see in Fig. 4(a), (10.7) is roughly periodic in $A_{\rm rel}$ with period $(\frac{1}{2}\,(2^B - 1))^{-1} = 0.0317$ and with shape function dictated by $|\zeta_{-\frac{1}{2}}(x)|^2$, see Subsec. 5.1. Also, the maximum and average THD^2-value agree with what is given by formulas (5.7–8). For the case that $M = 32$ we have that the second inequality in (8.10) holds, and, indeed, the THD^2 in Fig. 4(b) shows relatively modest variations as a function of $A_{\rm rel}$ at an average level correctly predicted by the right-hand side of (8.14). In Fig. 5 we have displayed (10.7) for the case (10.8) with $M = 16$ and extended range $A_{\rm rel} \in [0.5, 1.0]$. Now $2\pi M A_0/N$ is close to 1, whence neither of the two inequalities in (8.10) holds. We notice irregular behaviour and a hump in the values of (10.7) for $A_{\rm rel}$ close to

$$(10.9) \qquad \frac{N}{2\pi M} \cdot \frac{1}{\frac{1}{2}(2^B - 1)} = 0.64\ .$$

In Fig. 6 we show as a function of A_0

$$(10.10) \qquad 10 \log_{10} {\rm THD}^{2,sw}_{\rm rounding}$$

with

$$(10.11) \qquad N = 2048\ ; \qquad M = 2, 32\ ; \qquad B = 6\ .$$

When $M = 2$, we have that the first inequality in (8.15) holds, and we see in Fig. 6(a) that (10.10) is roughly periodic in $A_{\rm rel}$ with period $\frac{1}{2}\,(\frac{1}{2}\,(2^B - 1))^{-1} = 0.0159$ and with shape function dictated by $|\zeta_{-\frac{1}{2}}(x)|^2$, see Subsec. 6.1. This confirms formula (6.13), and also the maximum and average THD^2-value agree with what is given by formulas (6.14–15). Comparing Figs. 4(a), 6(a), we notice that (10.10) manifests itself at a level which is some 9 dB lower than the level at which (10.7) manifests itself while the periodicity interval for Fig. 6(a) is half of what it is in Fig. 4(a). When $M = 32$, we have that the second inequality in (8.15) holds, and we see in Fig. 6(b) that the values of (10.10) have become relatively insensitive to variations in A_0 while (8.16) correctly predicts the average values of (10.10). Notice that this average value is some 30 dB lower than the average value in Fig. 4(b). In Fig. 7 we have displayed (10.10) for the case (10.11) with $M = 4$ and extended range $A_{\rm rel} \in [0.5, 1.0]$. We notice irregular behaviour of (10.10) and a hump near $A_{\rm rel} = 0.64$ as in Fig. 5, however, at a level which is some 18 dB lower.

In Figs. 8, 9, 10 we illustrate the statement that wobbling with a ramp covering M sine periods yields qualitatively the same rounding THD^2 as in the non-wobbled case with a quantizer with $\log_2 M$ bits more. We have taken

$$(10.12) \qquad N = 16384\ ; \qquad M = 4 \quad ; \qquad B = 6 \text{ or } 8 \quad ,$$

(10.13) $N = 4096$; $M = 4$; $B = 6$ or 8 ,

(10.14) $N = 2048$; $M = 32$; $B = 6$ or 11 ,

where the first choice for B is for the case of wobbling. The three examples are for the respective cases in (8.10) and (8.15) that the first inequality, none of the inequalities, the second inequality holds.

We proceed by comparing the performance of the wobbling method and subtractive noise dithering method with respect to reducing the rounding THD^2 and stabilization of the SINAD^2. In Fig. 11 we show the THD^2 and SINAD^{-2} in dB for the ideal quantizer with input $f(t) = A_0 \sin 2\pi M t$ with

(10.15) $N = 2048$; $M = 16$; $B = 11$

and no wobbling (Fig. 11(a)), wobbling (Fig. 11(b)) and subtractive noise dithering (Fig. 11(c)) where in the latter case the dither noise has a uniform pdf over $[-\frac{1}{2}, \frac{1}{2}]$. The average THD^2-levels in Figs. 11(a), (b) are correctly predicted by (8.14) and (8.16) (second inequality in (8.10) and (8.15) holds here) at around -76 dB and -100 dB for $A_{\rm rel}$ close to 1. Also, the fluctuations of the SINAD^{-2} around the $\frac{1}{6}\,A_0^2$-law follow the formulas (8.19–20). In particular, we see that the fluctuations of the SINAD^{-2} in Fig. 11(b) have practically disappeared. From the noise dither pictures in Fig. 11(c) it is seen that the average THD^2 settles at a level of around -88 dB while the picture of the SINAD^{-2} is rather rough. We have also done simulations (not displayed here) with non-subtractive dither noise and with dither noise having a triangular pdf over $[-1, 1]$, but this just worsened the situation.

We finally demonstrate the potential of the wobbling method for retrieving a simulated nonlinearity h. We consider the case that N, M, B are as in (10.15) while h is given according to (10.3) through its coefficients

(10.16) $\alpha_5 = 5 \cdot 10^{-5}\,,\ \ \alpha_4 = 2.5 \cdot 10^{-5}\,,\ \ \alpha_3 = 1.5 \cdot 10^{-4}\,,\ \ \alpha_2 = 3.5 \cdot 10^{-5}\,,$
$\alpha_1 = 1\,,\ \ \alpha_0 = 0\,.$

In Fig. 12 we display for A_0 as in (10.2) the THD^2-values (dB) defined in (10.4–5), with the solid line the THD^2 due to h and the dashed-dotted line and the dotted line the THD^2 before and after wobbling. Clearly, the non-linearity h yields a distortion well below the distortion due to the rounding before wobbling and well above the distortion due to rounding after wobbling. In Fig. 13 we show the reconstruction results, case $A_{\rm rel} = 0.95$, where the formula (9.8) for $\tilde{h}^a$ is used with $\tilde{F}^a_{K,n}$ given in (9.7) for the cases (a) $a = nw$ (no wobbling), (b) $a = sw$ (wobbling as in Fig. 12). Clearly, we get a bad reconstruction result in Fig. 13(a) and a good one in Fig. 13(b). In Fig. 13(c) we show the reconstruction result for this case where now the coefficients $\tilde{F}^a_{K,n}$ for (9.8) are obtained from a subtractively dithered sine $A_0 \sin 2\pi M t$ (uniform pdf). The non-linearity h manifests itself through its THD^2-values at a level that noise dithering cannot accurately reconstruct h, see Figs. 11(c) and Fig. 12. Indeed, it is seen that the estimate in Fig. 13(c) is of considerably lower quantity than the one of Fig. 13(b), and this even gets worse for other forms of dithering like non-subtractive noise-dithering (uniform pdf) or using dither noise with a triangular pdf over $[-1, 1]$.

Acknowledgement

The authors want to thank their (ex, present,future) colleagues Mark Looijer, Ronald de Vries, Geert Seuren and Ram Pratap Aditham for their support, encouragement and actual contributions to the process of validating the wobbling method in theory and by simulations and measurements. They also want to thank Stan Baggen, for coining the term “wobbling” for the deterministic subtractive dithering technique studied in this paper, and Ronald Aarts, for useful discussions on applications of the method not necessarily limited to the context of testing.

References

[1] M. Abramowitz et al., Handbook of Mathematical Functions, Dover, New York, 1970.

[2] M.T. Abuelma’atti, Spectrum of a nonlinearly quantized sinusoid, IEE Proc. G 136 (6) (December 1989) 313–316.

[3] P. Carbone, D. Petri, Effect of additive dither on the resolution of ideal quantizers, IEEE Trans. Instrum. Meas. 43 (3) (June 1994) 389–396.

[4] N. Csizmadia, A.J.E.M. Janssen, Estimating the integral non-linearity of AD-converters via the frequency domain, preprint.

[5] T.A.C.M. Claasen, A. Jongepier, Model for the power spectral density of quantization noise, IEEE Trans. Acoust. Speech, Signal Processing 29 (4) (August 1981) 914–917.

[6] R.M. Gray, T.G. Stockham, Dithered quantizers, IEEE Trans. Information Theory 39 (3) (May 1993) 805-812.

[7] A.J.E.M. Janssen, Fourier Analysis and Synthesis for a Mildly Non-Linear Quantizer, Nat. Lab. Unclassified Report 801/99, Philips Research Laboratories Eindhoven, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands; also downloadable from http://www.research.philips.com/manuscript/search.html

[8] I. De Lotto, G.E. Paglia, Dithering improves A/D converter linearity, IEEE Trans. Instrum. Meas. 35 (2) (June 1986) 170–177.

[9] J. Vanderkooy, S.P. Lipshitz, Dither in digital audio, J. Audio Eng. Soc. 35 (12) (December 1987) 966–975.

[10] R. de Vries, A.J.E.M. Janssen, Decreasing the sensitivity of ADC test parameters by means of wobbling, Proc. 16^{th} IEEE VLSI test symposium, Monterey, CA, 27–30 April 1998, pp. 386–391.

[11] E.T. Whittaker, G.N. Watson, Modern Analysis, Cambridge-at-the-University Press, Cambridge, 1962.

Figure captions

Fig. 1. Plot of the Hurwitz zeta functions $\zeta_{-\frac{1}{2}}, \zeta_{-\frac{3}{2}}$, see (5.4), (5.19) (solid line, dashed line, respectively).

Fig. 2. Plot of $2\pi^2 m^2\, S_m(n)$ and $2\pi^2 m^2\, \overline{S_m(n)}$, see (5.10) and (5.12), as a function of n for $m = 3$, $A_0 = 32$ (dotted line, solid line, respectively).

Fig. 3. Plot of $\overline{S(n)}$, see (5.15), as a function of n for $A_0 = 32$ in dB.

Fig. 4. Plot of THD^2 due to rounding, non-wobbled, in dB for $N = 2048$, $B = 6$ and (a) $M = 2$, (b) $M = 32$ as a function of $A_0 = \frac{1}{2}\, A_{\mathrm{rel}}(2^B - 1)$, $A_{\mathrm{rel}} \in [0.8, 1.0]$.

Fig. 5. Plot of THD^2 due to rounding, non-wobbled, in dB for $N = 2048$, $B = 6$ and $M = 16$ as a function of $A_0 = \frac{1}{2}\, A_{\mathrm{rel}}(2^B - 1)$, $A_{\mathrm{rel}} \in [0.5, 1.0]$.

Fig. 6. Plot of THD^2 due to rounding, wobbled, in dB for $N = 2048$, $B = 6$ and (a) $M = 2$, (b) $M = 32$ as a function of $A_0 = \frac{1}{2}\, A_{\mathrm{rel}}(2^B - 1)$, $A_{\mathrm{rel}} \in [0.8, 1.0]$.

Fig. 7. Plot of THD^2 due to rounding, wobbled, in dB for $N = 2048$, $B = 6$ and $M = 4$ as a function of $A_0 = \frac{1}{2}\, A_{\mathrm{rel}}(2^B - 1)$, $A_{\mathrm{rel}} \in [0.5, 1.0]$.

Fig. 8. Plot of THD^2 due to rounding in dB for $N = 16384$, $M = 4$ and (a) $B = 6$, wobbled, (b) $B = 8$, non-wobbled.

Fig. 9. Plot of THD^2 due to rounding in dB for $N = 4096$, $M = 4$ and (a) $B = 6$, (b) $B = 8$, non-wobbled.

Fig. 10. Plot of THD^2 due to rounding in dB for $N = 2048$, $M = 32$ and (a) $B = 6$, (b) $B = 11$, non-wobbled.

Fig. 11. Plot of THD^2 and SINAD^{-2} in dB for the ideal quantizer with input $f(t) = A_0 \sin 2\pi M t$ where $N = 2048$, $B = 11$, $A_{\mathrm{rel}} \in [0.8, 1.0]$ and $M = 16$ with (a) no wobbling, (b) wobbling, (c) subtractive noise dithering using dither noise with a uniform pdf on $[-\frac{1}{2}, \frac{1}{2}]$.

Fig. 12. Plot of THD^2 in dB due to non-linearity h (solid line), due to rounding before wobbling (dashed-dotted line) and after wobbling (dotted line) for $N = 2048$, $M = 16$, $B = 11$ and non-linearity h as in (10.3) with $\alpha_5 = 5 \cdot 10^{-5}$, $\alpha_4 = 2.5 \cdot 10^{-5}$, $\alpha_3 = 1.5 \cdot 10^{-4}$, $\alpha_2 = 3.5 \cdot 10^{-5}$, $\alpha_1 = 1$, $\alpha_0 = 0$ as a function of $A_0 = \frac{1}{2}\, A_{\mathrm{rel}}(2^B - 1)$, $A_{\mathrm{rel}} \in [0.8, 1.0]$.

Fig. 13. Reconstruction results of the non-linear part $h(x) - x$ (solid line) for $|x| \le 0.95 \cdot \frac{1}{2}\,(2^B - 1)$ with h as in Fig. 12 and $\tilde{h}(x)$ given by (9.8) with $\tilde{F}^{a}_{K,n}$ given in (9.7) (a) non-wobbled, (b) wobbled as in Fig. 12, (c) dithered using subtractive dither noise with a uniform pdf on $[-\frac{1}{2}, \frac{1}{2}]$ (dashed-dotted lines).

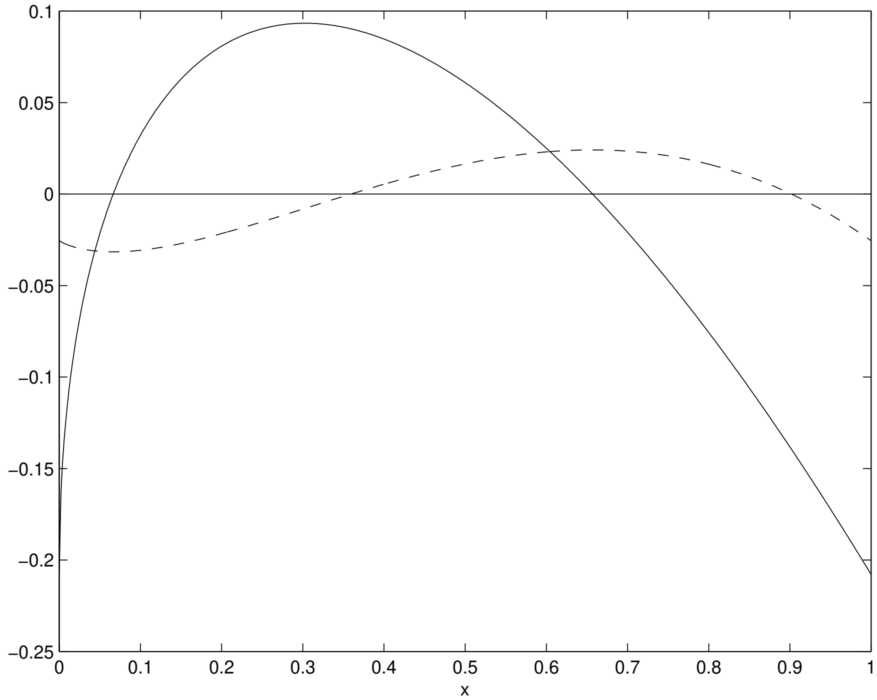

Fig. 1

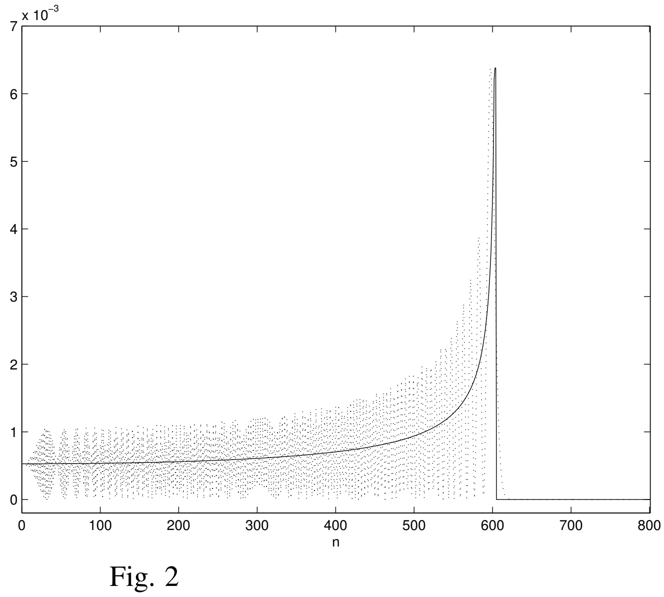

Fig. 2

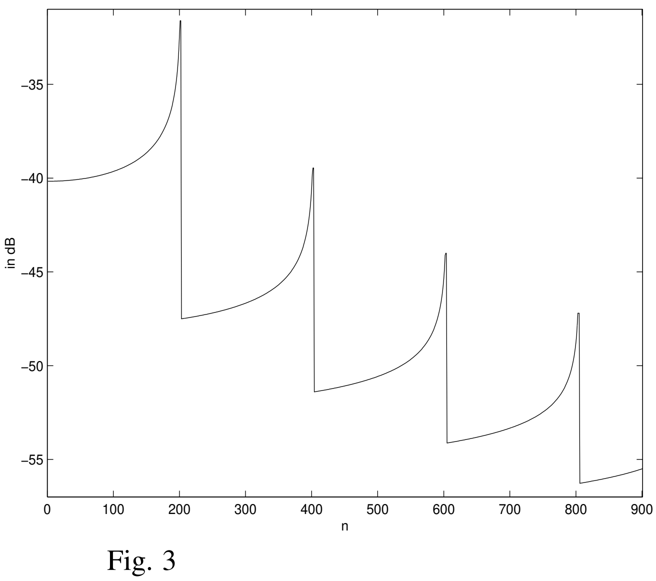

Fig. 3

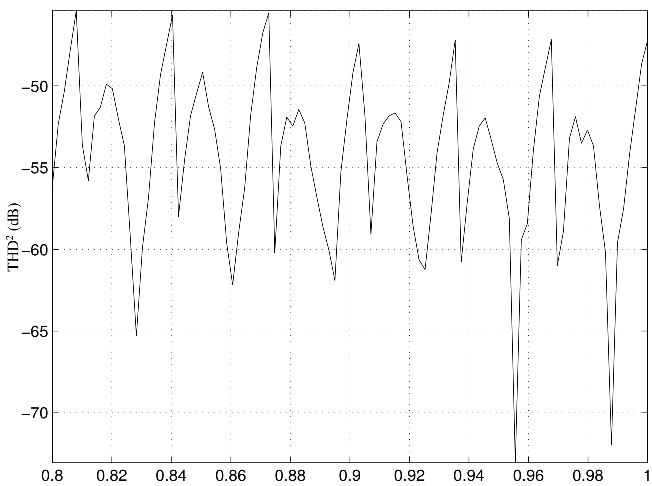

Fig. 4a – relative amplitude

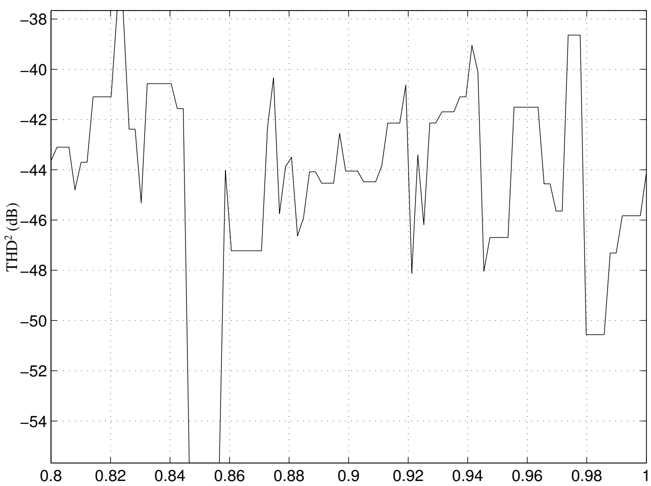

Fig. 4b – relative amplitude

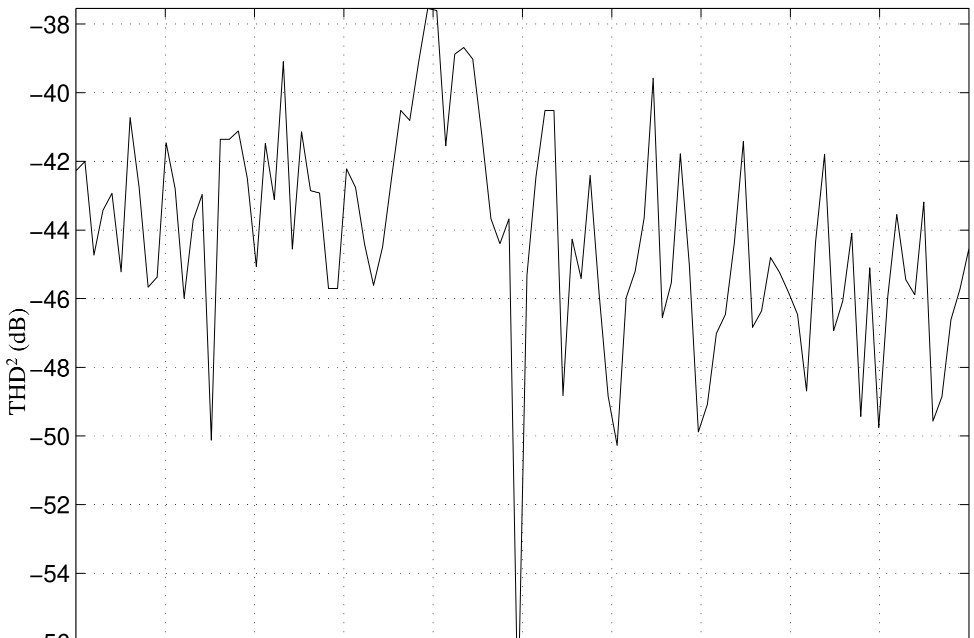

Fig. 5 – relative amplitude

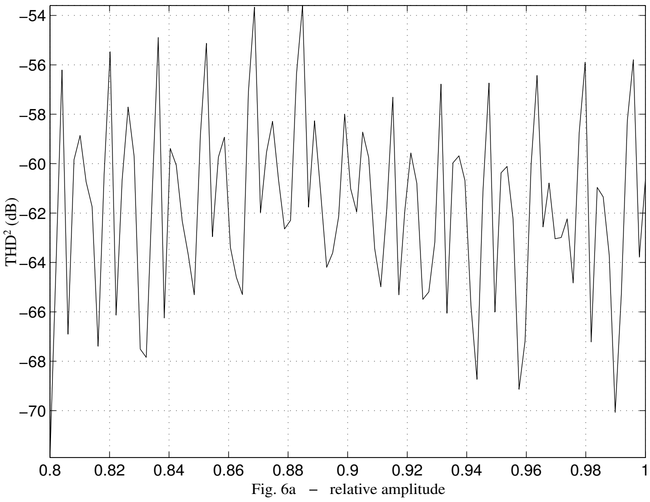

Fig. 6a – relative amplitude

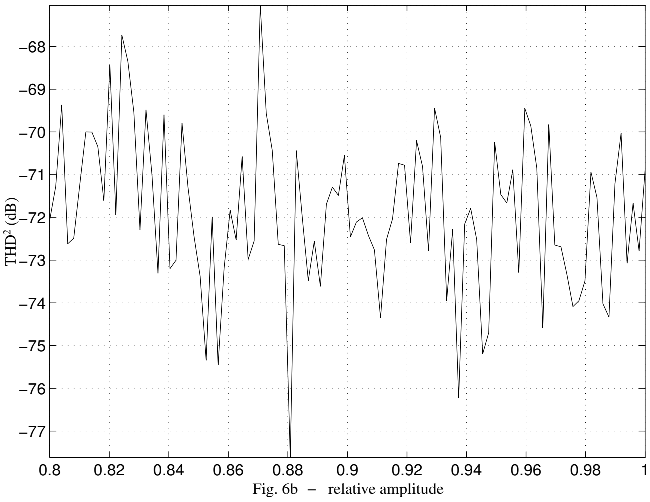

Fig. 6b – relative amplitude

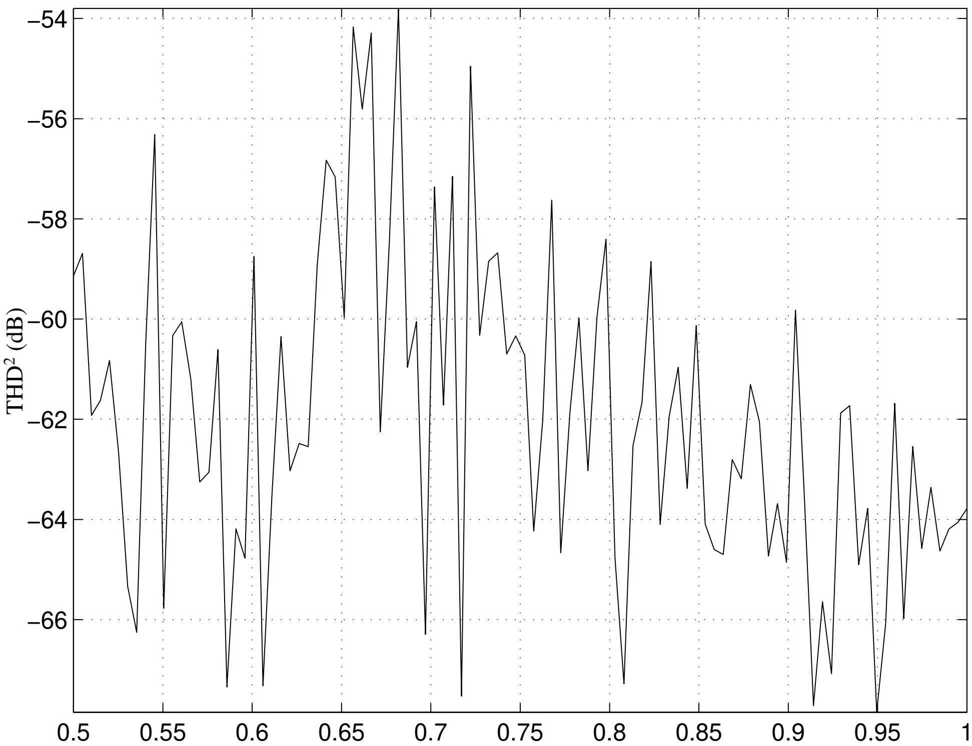

Fig. 7 – relative amplitude

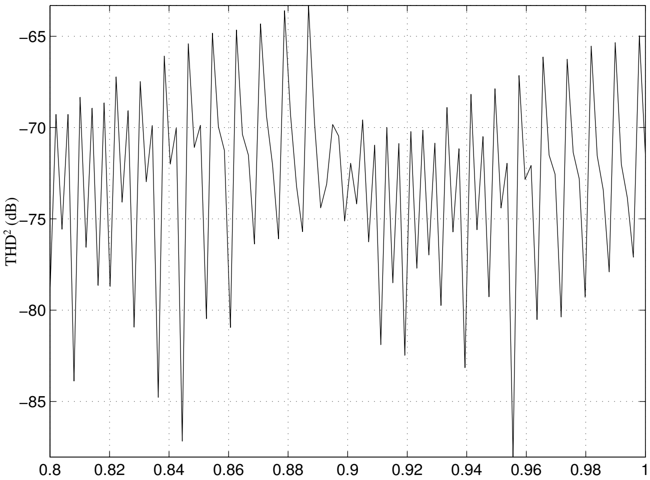

Fig. 8a – relative amplitude

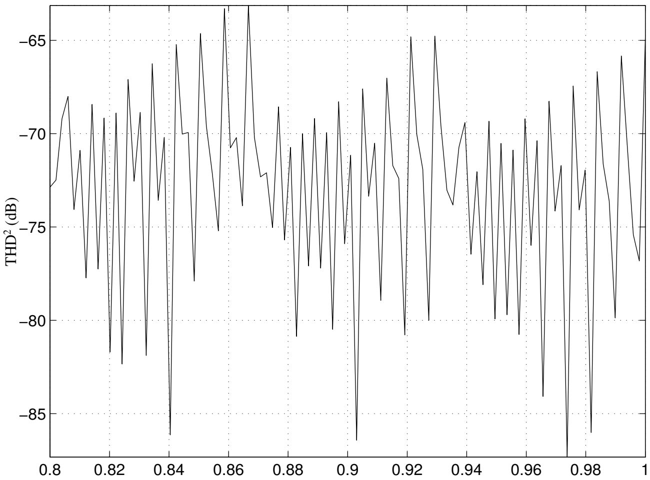

Fig. 8b – relative amplitude

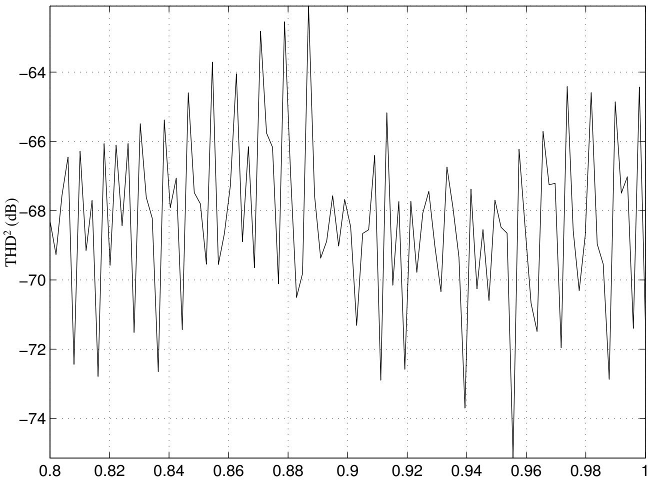

Fig. 9a – relative amplitude

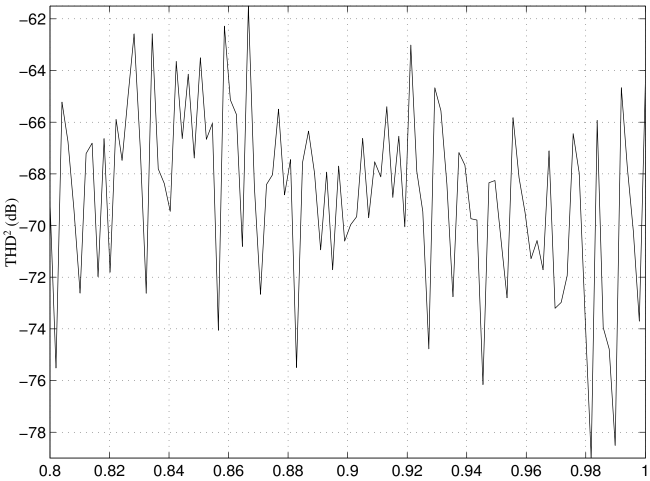

Fig. 9b – relative amplitude

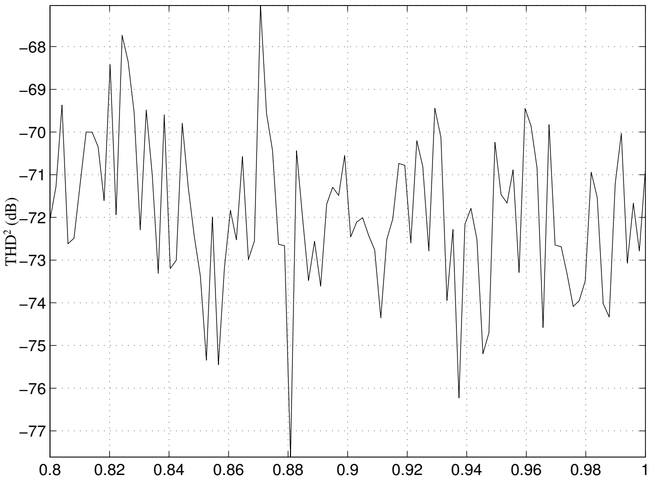

Fig. 10a – relative amplitude

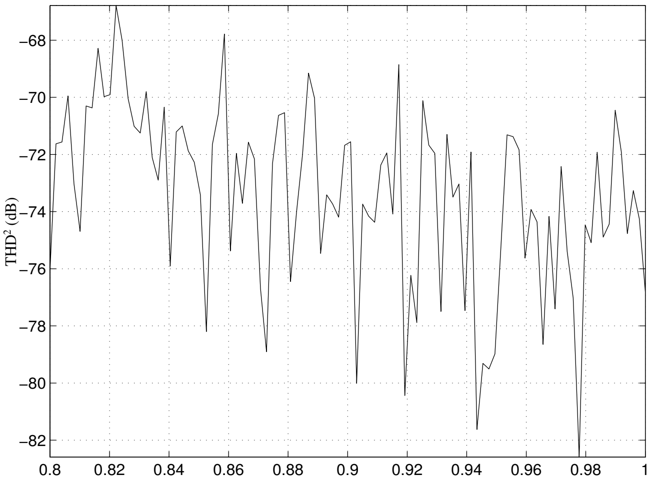

Fig. 10b – relative amplitude

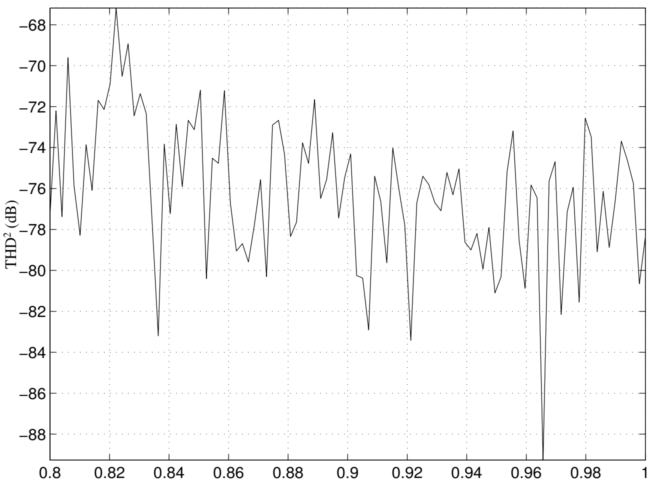

Fig. 11a – relative amplitude

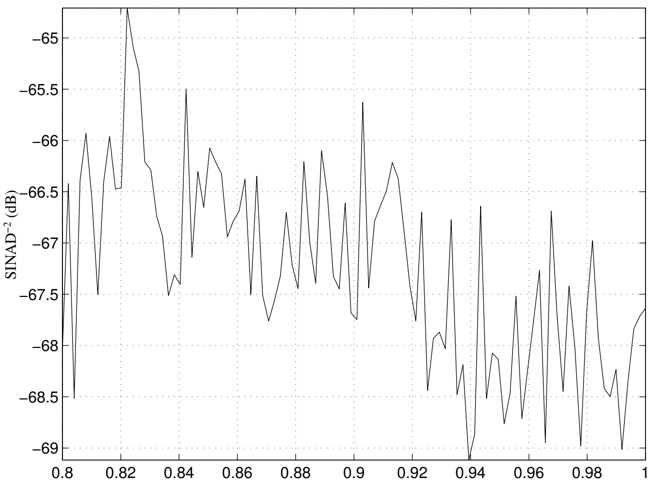

Fig. 11a – relative amplitude

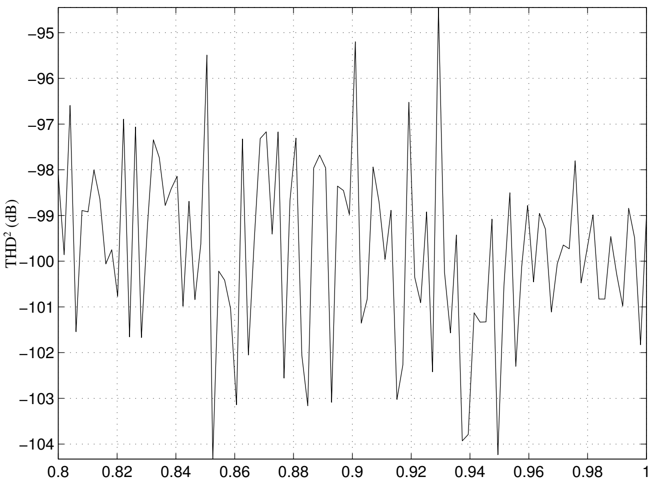

Fig. 11b – relative amplitude

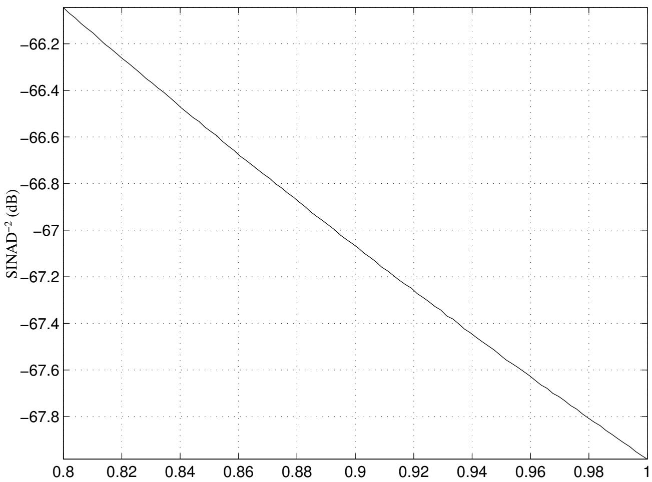

Fig. 11b – relative amplitude

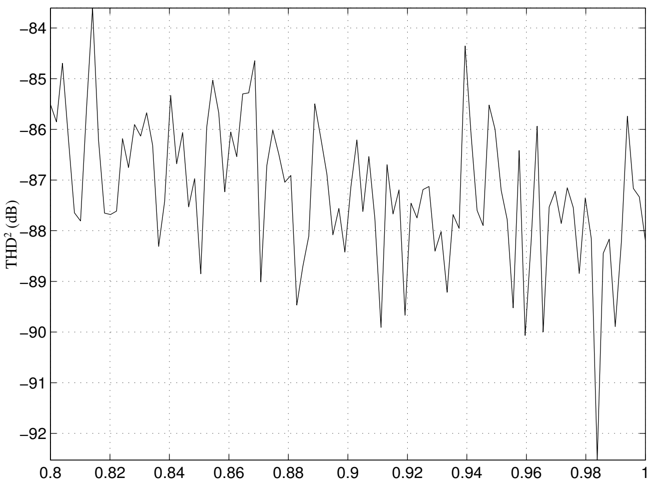

Fig. 11c – relative amplitude

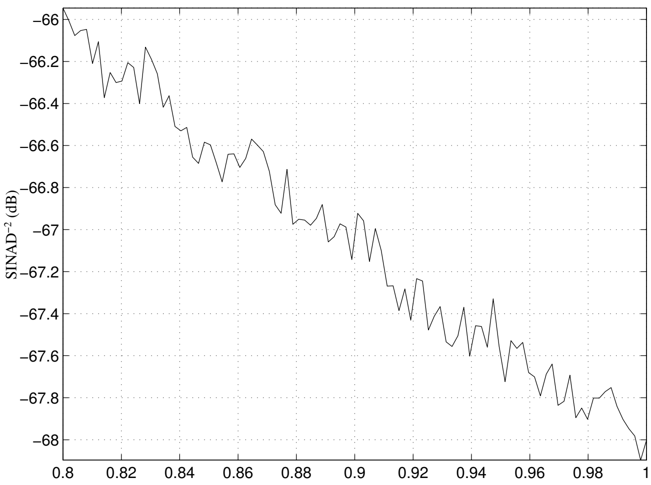

Fig. 11c – relative amplitude

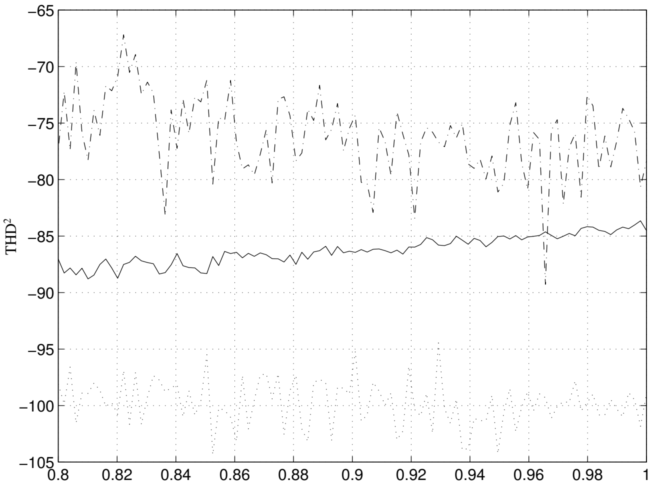

Fig. 12 – relative amplitude

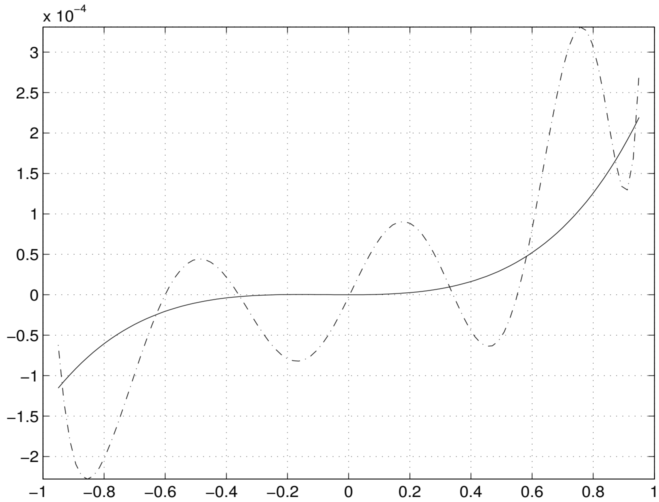

Fig. 13a

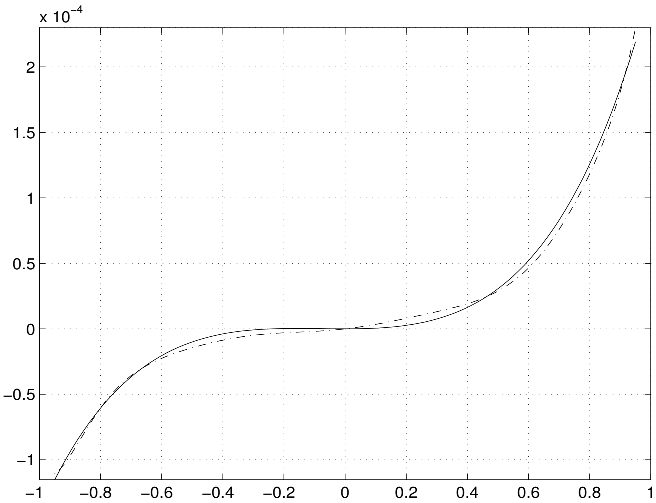

Fig. 13b

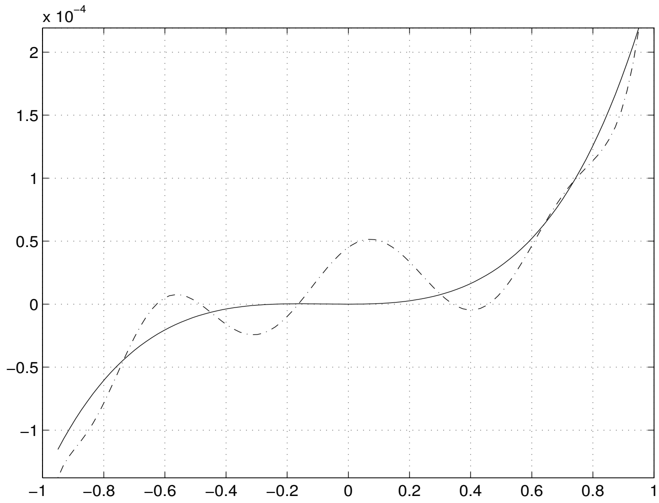

Fig. 13c