Dithering and data compression

Schobben, D.W.E.

Award date:
1995

EINDHOVEN UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING

Dithering and Data compression

by D.W.E. Schobben

Masters thesis
Period of work: January 2, 1995 - July 31, 1995
under supervision of:
Ir. R.A. Beuker and Ir. A.W.J. Oomen (Philips Research Laboratories)
Ir. J.H.F. Ritzerfeld and Dr. Ir. P.C.W. Sommen (Eindhoven University)

Abstract

In many analyses, quantisation errors are modelled as being input-signal independent, uniformly distributed white noise. This approximation fails in the case that the amplitude of the input signal is comparable to the quantiser stepsize.

Non-subtractive dithering is a technique which is capable of rendering a number of quantisation error moments functionally independent of the input signal, at the expense of an increase of the error noise floor. Subtractive dithering is a technique which guarantees the total functional independency of the quantising error with respect to the signal input.

A drawback of the two mentioned techniques is that the entropy of the quantised signal increases compared to the case that no dither signal is applied. This thesis provides an extensive entropy analysis, applied to the various dither signals.

Experimental results indicate that the coding efficiency of a PCM system can be significantly improved using dither, in both audio and video applications. Using dither to quantise DCT coefficients in a video system did not improve the perceptual quality. In a four band audio subband system however, improvements are reported when dithering the first subband. The noise shaping technique can be effectively used to shape the resulting white noise to less perceptible frequencies.

Contents

Chapter 1

Introduction

The amount of digital information to be transmitted and stored increases at a phenomenal rate. As a result, the use of efficient data compression techniques has become crucial. Data compression can be divided into lossy and lossless compression. In a lossless data compression system, the recovered data is identical to the original data. In audio and video compression systems lossy compression suffices. The goal in a lossy compression system is to generate a compressed-decompressed sequence that perceptually resembles the original sequence as good as possible at a the lowest possible bit rate. Compression ratios are much higher for lossy compression than for lossless compression. The data compression techniques described in this thesis are all lossy.

1.1 Source Coding

In a source coding system, a time-discrete signal is transformed, quantised and lossless-coded prior to transmission or storage [1], as depicted in Figure 1.1. At the receiver side, the signal

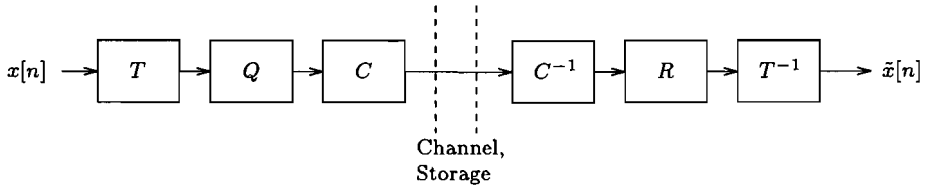


Figure 1.1: Blocks of a source coding system

is decoded, reconstructed and inversely transformed. The transformation decorrelates the signal, the quantising operation removes the irrelevance, and the lossless coder removes the remaining redundancy of the signal. Only transmission over a noiseless channel is considered. This model can easily be extended to transmission over a noisy channel by using error correcting codes.

In many analyses, quantisation errors are modelled as being input-signal independent, uniformly distributed white noise. This approximation fails in the case that the amplitude of the

input signal is comparable to the quantiser stepsize. Quantisation always introduces rounding errors. Quantising errors in undithered systems are not random. They are deterministically related to the input of the quantiser. In general, quantisation errors are not uniformly distributed either. The signal dependency can be observed as noise modulation and distortion. Note that time sampling never introduces errors, given that the signal is properly band-limited and Nyquist sampling theorem is met.

Non-subtractive dithering is a technique which is capable of rendering a number of quantisation error moments functionally independent of the input signal, at the expense of an increase of the error noise floor. Subtractive dithering is a technique which guarantees the functional independency of the quantising error with respect to the signal input.

A survey of the theorems concerning the uniform distribution, functional independency and statistical independency of the quantisation error with respect to the input signal in undithered, subtractively dithered and non-subtractively dithered systems can be found in [2], and is summarised in Appendix A. The abbreviations used in this appendix are explained in the list of symbols and abbreviations on page 57.

In the following sections, an introduction to undithered, subtractively dithered and non-subtractively dithered quantisation will be presented. The quantisers used are assumed to be uniform and infinite. This will be discussed in more detail in Section 1.5. The theorems mentioned can be found in Appendix A.

1.2 Undithered quantisation

In the classical model of undithered quantisation, the quantisation operation is modelled by the addition of white noise with a Rectangular Probability Density Function, RPDF noise η, of power $\frac{\Delta^2}{12}$, as depicted in Figure 1.2. This approximation only holds if the quantisation

Figure 1.2: Classical model undithered quantiser

stepsize Δ is small compared to the width of the pdf of the input signal, w. A statistical model valid for inputs with arbitrary statistical properties was first developed by Widrow [3]. When examining the distortion and the quantisation noise modulation introduced by the non-linear quantiser, the joint pdf of the error signal must be considered, since it determines the power spectral density (PSD) of the error signal. In an undithered quantising system, the joint characteristic function, i.e. the Fourier transformed joint probability density function, must obey Theorem 2 in Appendix A, in order to guarantee the uniform distribution of the quantisation error samples separated in time. It is clear that in general this is not true for practical signals. Note that the uniform distribution of the quantisation error does not imply

that the quantisation error is independent of the system input.

1.3 Subtractive dither

Dithering means to tremble or to vibrate. This trembling refers to the quantisation process, which has fixed decision and representation levels in the undithered case. Dithering the quantiser has the effect of randomising the quantisation error. In order to obtain this effect, the dither must have a random nature. It can be perceptually disturbing when the quantising error, $\epsilon = x - \tilde{x}$, is input signal dependent. The first use of subtractive dither must be credited to Roberts [4], who applied it to picture coding. A theoretical investigation of subtractive dither was done by Schuchman [5], a student of Widrow. Schuchman derived a condition on the dither pdf which guarantees the statistical independency of the quantisation error with respect to the input signal. Later, a second order analysis was made by Sherwood [6]. In a subtractively dithered-quantising system, the dither signal is added prior to quantisation, and subtracted after transmission, as is depicted in Figure 1.3. A subtractively dithered quantiser can be considered as a quantiser with randomised decision and representation levels. If the

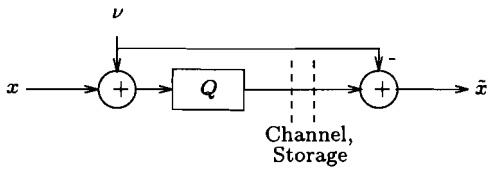


Figure 1.3: Subtractively dithered quantiser

dither is chosen conform Theorem 5, all error moments are independent of the input signal, and are of the sort postulated in the classical model. Subtractive dithering not only renders all error moments independent of the system input, it also renders the total error functionally independent of the input signal. There is no noise penalty and the total noise power remains $\frac{\Delta^2}{12}$. A disadvantage of subtractive dither is that the dither is needed at the receiver. The dither is, however, usually not transmitted over the channel, but is generated algorithmically at both the transmitter and the receiver side. Synchronisation of the dither can be done for example by using the first sample to be transmitted as a seed for the random number generator that produces the dither. Note that the entropy of the quantised signal in a subtractively dithered system is identical to that of a non-subtractively dithered system, when using the same dither. The quantiser stepsize, Δ, is commonly referred to as an LSB ("Least Significant Bit"), since a change in input signal level of one step height corresponds to a change in the LSB of binary coded output. The most common dither that obeys Theorem 5, is white noise with a rectangular pdf, RPDF, of 1 LSB amplitude.

Dither, either subtractive or non-subtractive linearises the expectation of the transfer function of the quantisation scheme. In the undithered case the expectation of the transfer function is identical to the staircase quantiser transfer function. This is depicted in the first row of Figure 1.6. In the third row of the figure, RPDF dither of 1 LSB top-to-top amplitude is used,

and the expectation of the quantiser output equals the quantiser input, so that the quantiser transfer function is linearised. The second row shows the expectation of the quantiser transfer function when RDPDF dither of $\alpha\Delta$ amplitude is used. The transfer function is only partially linearised. Again, the transfer function is linearised if α is an integer.

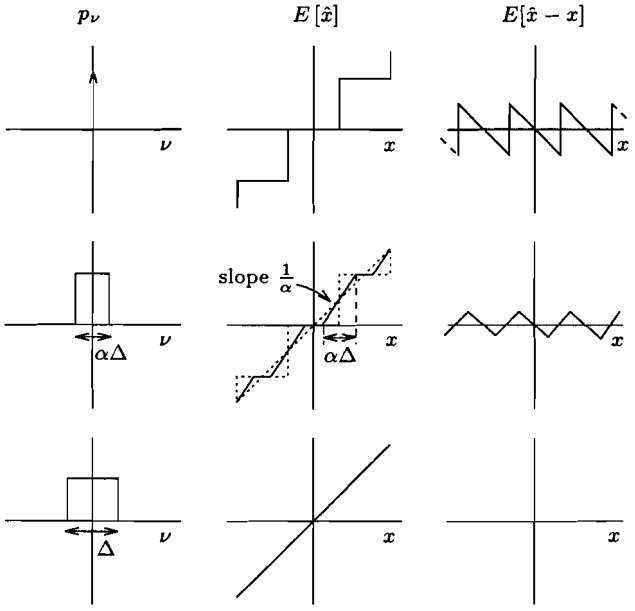


Figure 1.4: Quantiser linearisation by means of RPDF dither

1.4 Non-subtractive dither

In the case of non-subtractive dithering, the dither signal, ν, is added to the input signal prior to quantisation and not subtracted afterwards, as is depicted in Figure 1.5. Non-subtractive

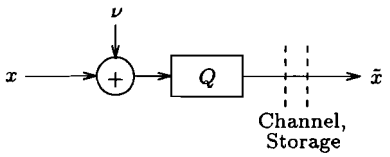


Figure 1.5: Non-subtractively dithered quantiser

dither is of interest since it can be incorporated in existing audio and video systems, and it does not have the problem of regenerating and synchronising the dither. In a non-subtractively dithered quantising system it is not possible to render the quantising error white, but it is possible to render some of the statistical properties of the quantising error independent of

the input. If the dither is chosen conform Theorem 6, $E[\epsilon^m|x]$ is functionally independent of x. Note that if this theorem is satisfied for $m = P$, this does not guarantee that it is also satisfied for the p^{th} error moment, $p \leq P$. It of interest to render the first N moments of the total error functionally independent of the system input. This is the case if Theorem 7 is satisfied. Clearly, the only class of white dither which renders the first N moments of the error independent of the system input is a summation of N independent rectangular-pdf white noise sources, each of 1 LSB amplitude, as stated by Theorem 8. The resulting total error power will be $\frac{1+N}{12}\Delta^2$, which corresponds with an increase of $\frac{N}{12}\Delta^2$ in noise power. Besides this noise penalty, the entropy of the quantised signal increases as a consequence of the addition of dither. Since only the first two moments are believed to be audible, there is no point in rendering higher moments than the first and the second independent of the system input for audio applications. For video, the third moment can be seen, but is of significantly less importance than the first two error moments. Audio and video experiments indicate that dither with a triangular-pdf, TPDF dither, yields the perceptually least annoying total error signal. TPDF dither is generated by the summation of two independent RPDF signals.

1.5 Quantisers

A quantiser is defined as a device with input x and output $Q(x)$, which maps input intervals on discrete output values, i.e.

$$\forall x : \ \alpha_{i-1} < x \leq \alpha_i \Rightarrow Q(x) = \beta_i \, . \tag{1.1}$$

The α_i denote the decision levels, the β_i denote the representation levels and $i \in \mathbb{Z}$. Without loss of generality it is stated that

$$\forall i : \ \alpha_i > \alpha_{i-1}, \ \beta_i > \beta_{i-1} \, . \tag{1.2}$$

The quantiser error signal ϵ is given by

$$\epsilon(x) = Q(x) - x \, . \tag{1.3}$$

Throughout the sequel, quantisers discussed will be assumed to be uniform and infinite, so that the input signal is never clipped. The decision levels and the representation levels are then given by

$$\forall i : \ \alpha_i = \left(i + \tfrac{1}{2}\right)\Delta + \kappa, \ \beta_i = i\Delta + \kappa \, , \tag{1.4}$$

where κ is a constant, $-\frac{\Delta}{2} < \kappa \leq \frac{\Delta}{2}$. Because α_i and β_i obey (1.4), the quantising error never exceeds $\frac{\Delta}{2}$, i.e.

$$-\frac{\Delta}{2} < \epsilon \leq \frac{\Delta}{2} \, . \tag{1.5}$$

The output of the quantiser is depicted in Figure 1.6, and is given by

$$Q(x) = \Delta \left\lfloor \frac{x - \kappa}{\Delta} + \frac{1}{2} \right\rfloor + \kappa \, , \tag{1.6}$$

where $\lfloor \cdot \rfloor$ indicates the floor operator, which returns the greatest integer less or equal to its argument. Note that in the case κ equals zero, the quantiser has a properly defined zero

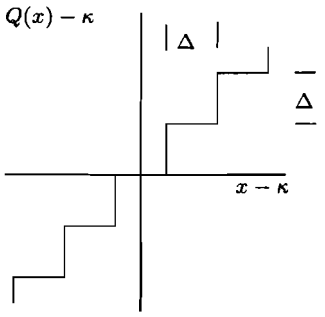


Figure 1.6: General quantiser characteristic

output level and is generally referred to as a mid-tread quantiser. The output of this quantiser is given by

$$Q_{MT}(x) = \Delta \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor . \tag{1.7}$$

The quantiser is of the mid-riser type in the case that κ equals $\frac{\Delta}{2}$. The output of this quantiser is given by

$$Q_{MR}(x) = \Delta \left\lfloor \frac{x}{\Delta} \right\rfloor + \frac{\Delta}{2} . \tag{1.8}$$

Chapter 2

Rate-Distortion for uniform sources

When an audio or video source is encoded, it is useful to model the source by means of a probability density function. A video signal is roughly estimated by a uniform distribution and an audio sequence can be modelled by a Laplacian distribution. In this chapter, the statistics of the output of a uniform quantiser are calculated for an input signal with a uniform pdf, generated by a memoryless source, i.e.

$$p_x(x) = \Pi_w(x)\,, \tag{2.1}$$

where the probability density function of x is denoted as $p_x(x)$, and $\Pi_w(x)$ denotes the rectangular window function, defined as

$$\Pi_w(x) \triangleq \begin{cases} \frac{1}{w}, & -\frac{w}{2} < x \leq \frac{w}{2} \\ 0, & \text{otherwise} \end{cases} . \tag{2.2}$$

In this chapter, w is referred to a the input signals pdf width. All results have obvious analogues for non-uniform input distributions, although entropy and distortion curves will be different.

2.1 Undithered Distortion

A quantity of great interest is the average distortion that is introduced by the quantiser. This distortion is investigated for a input signal with a uniform input pdf. The r^{th} power distortion, $E[\epsilon^r]$, also know as the r^{th} error moment, is defined as

$$\begin{aligned} E[\epsilon^r] &= E[(Q(x) - x)^r]\,, \\ &= \sum_i \int_{\alpha_{i-1}}^{\alpha_i} (\beta_i - x)^r p_x(x) dx\,, \\ &= \sum_{i=-\infty}^{\infty} \int_{\kappa+(i-\frac{1}{2})\Delta}^{\kappa+(i+\frac{1}{2})\Delta} (\kappa + i\Delta - x)^r p_x(x) dx\,. \end{aligned} \tag{2.3}$$

The most commonly used error criterion, is the Mean Square Error ($r = 2$). Note that a lower MSE does not necessarily imply a higher subjective reconstructed quality. The error

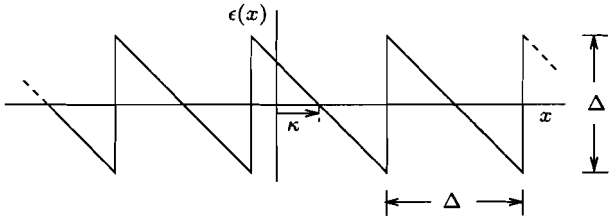


Figure 2.1: quantisation error

moments for an undithered input signal with a uniform pdf follows from inserting (2.1) in (2.3)

$$E[\epsilon^r] = \frac{1}{w}\int_{-\frac{w}{2}}^{\kappa+(i_L+\frac{1}{2})\Delta}(\kappa+i_L\Delta-x)^r dx + \frac{1}{w}\sum_{i=i_L+1}^{i_R-1}\int_{\kappa+(i-\frac{1}{2})\Delta}^{\kappa+(i+\frac{1}{2})\Delta}(\kappa+i\Delta-x)^r dx + \frac{1}{w}\int_{\kappa+(i_R-\frac{1}{2})\Delta}^{\frac{w}{2}}(\kappa+i_R\Delta-x)^r dx . \quad (2.4)$$

With

$$\begin{aligned} i_L &= -\left\lfloor \frac{\frac{w}{2}+\kappa}{\Delta}+\frac{1}{2}\right\rfloor , \\ i_R &= \left\lfloor \frac{\frac{w}{2}-\kappa}{\Delta}+\frac{1}{2}\right\rfloor . \end{aligned} \quad (2.5)$$

The integers i_L and i_R represent the indices corresponding with the first respectively the last non-zero term from (2.3). The error signal $\epsilon(x)$ is depicted as a function of the quantiser input x in Figure 2.1. It appears from this figure that all terms in between i_L and i_R add the same contribution, so that the integrals of (2.4) can be calculated analytically

$$E[\epsilon^r] = \frac{-1}{w(r+1)}\left[(\kappa+i_L-x)^{r+1}\Big|_{-\frac{w}{2}}^{\kappa+(i_L+\frac{1}{2})\Delta} + (-x)^{r+1}\Big|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}(i_R-i_L-1) + (\kappa+i_R\Delta-x)^{r+1}\Big|_{\kappa+(i_R-\frac{1}{2})\Delta}^{\frac{w}{2}}\right] . \quad (2.6)$$

Further simplification yields

$$E[\epsilon^r] = \begin{cases} \frac{1}{w(1+r)}(\xi_L-\xi_R), & \text{for r odd} \\ \frac{1}{w(1+r)}(\xi_R+\xi_M+\xi_L), & \text{for r even} \end{cases} . \quad (2.7)$$

With

$$\begin{aligned} \xi_L &= (\tfrac{w}{2}+\kappa+i_L\Delta)^{r+1} , \\ \xi_R &= (\tfrac{w}{2}-\kappa-i_R\Delta)^{r+1} , \\ \xi_M &= 2(i_R-i_L)(\tfrac{\Delta}{2})^{r+1} . \end{aligned} \quad (2.8)$$

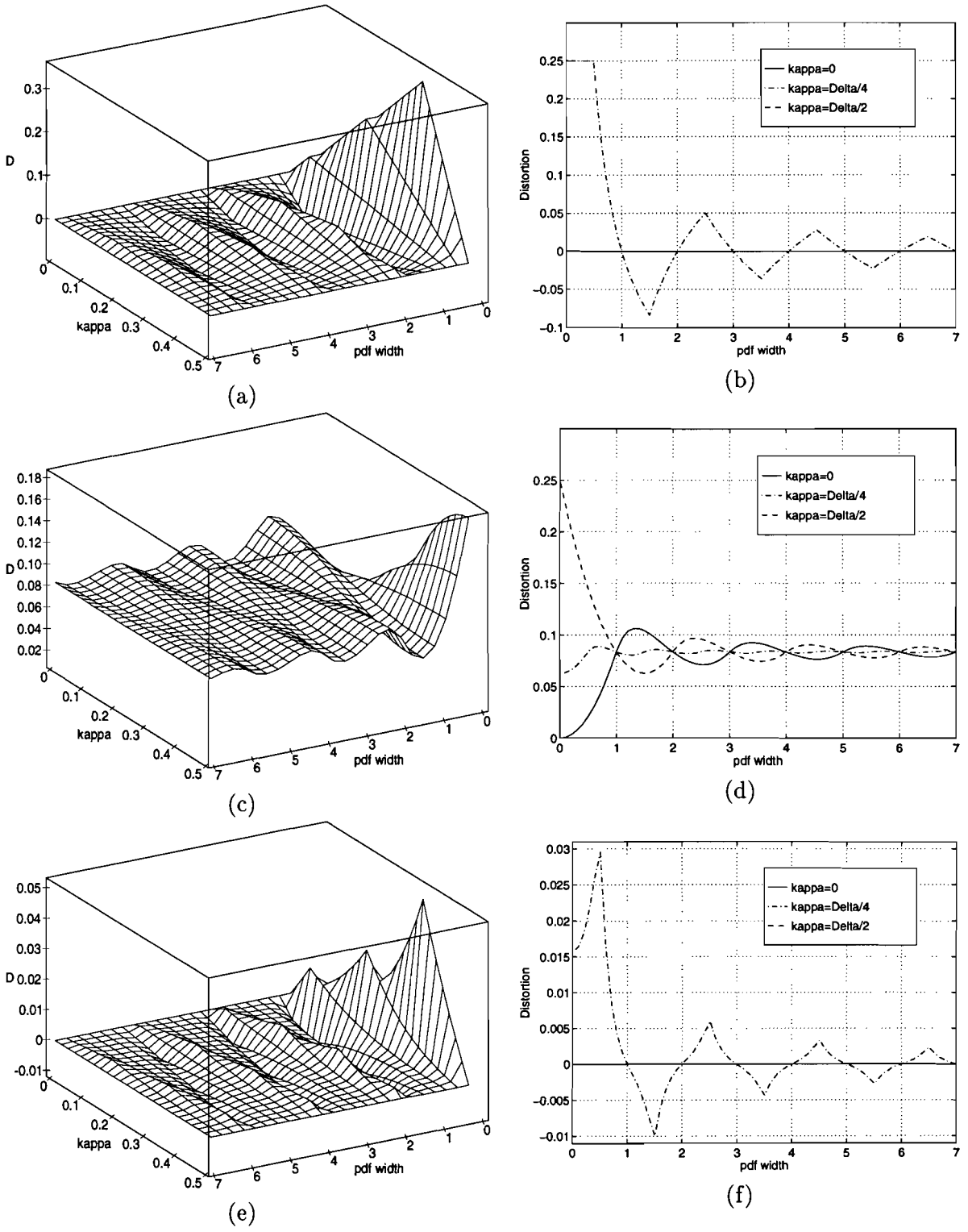


Figure 2.2: First, second and third-order distortion of an uniform quantiser with a stepsize $\Delta = 1$ and an uniform input pdf are shown in (a), (c) and (e) respectively. The same plots are included in (b), (d) and (f), for κ=0,$\frac{\Delta}{4}$ and $\frac{\Delta}{2}$, in order to ease the comparison.

Note that (2.7) has symmetrical properties, since

$$\begin{aligned} i_L|_{-\kappa} &= -i_R|_{\kappa}\,, \\ \xi_L|_{\kappa} &= -\xi_R|_{\kappa}\,, \\ \xi_M|_{-\kappa} &= -\xi_M|_{\kappa}\,, \end{aligned} \tag{2.9}$$

from which it follows that

$$E[\epsilon^r]|_{\kappa} = \begin{cases} E[\epsilon^r]|_{-\kappa}, & \text{for r odd} \\ E[\epsilon^r]|_{\kappa}, & \text{for r even} \end{cases}. \tag{2.10}$$

It follows from (2.7) that $E[\epsilon^r] = 0$ in the case r is odd and κ equals zero or $\frac{\Delta}{2}$. Furthermore, the zeroth moment is equal to unity, for all κ. The choice of κ is especially important for input signals with small pdfs with regard to the stepsize Δ. The choice of κ has no influence on the r^{th} power distortion for large w with respect to Δ, as follows from

$$E[\epsilon^r]|_{w \gg \Delta} = \begin{cases} 0, & \text{for r odd} \\ \frac{1}{(1+r)}(\frac{\Delta}{2})^r, & \text{for r even} \end{cases}. \tag{2.11}$$

These error moments agree with the moments that follow from the classical model, i.e.

$$E[\epsilon^r] = \frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \epsilon^r d\epsilon = \begin{cases} 0, & \text{for r odd} \\ \frac{1}{(1+r)}(\frac{\Delta}{2})^r, & \text{for r even} \end{cases}. \tag{2.12}$$

Note that only the error distribution resemblances the classical model. The error itself is still deterministically related to the quantiser input, instead of being independent white noise as assumed in the classical model. Since the uniform pdf modelling especially applies to video, the first, the second error and also the third error moment are plotted for $0 \leq \kappa \leq \frac{\Delta}{2}$ and $\Delta = 1$ in Figure 2.2. Note that no loss of generality occurs, when $\Delta = 1$, since only the choice of Δ relative to w is of importance. The simulations confirm that $E[\epsilon]$ and $E[\epsilon^3]$ are zero for mid-tread and mid-riser quantisation. Furthermore $E[\epsilon^3]$ is neglectably small for all κ, compared to the first and the second error moment. The error moments converge fast to the values stated by the classical model, for input pdf widths exceeding the quantiser stepsize Δ.

2.2 Undithered Entropy

After quantisation, the discrete signal is lossless coded prior to transmission or storage. The lower bound for encoding a source output into binary data is called the entropy of the signal and is determined by the self-information of all representation levels [9]. The probability of the quantiser output equaling level β_i is given by

$$p_i = p(Q(x) = \beta_i) = \int_{\alpha_{i-1}}^{\alpha_i} p_x(x)dx\,. \tag{2.13}$$

The self-information of β_i, in bits, is given by

$$I_i = -\log_2(p_i)\,. \tag{2.14}$$

In order to ease the notation of the entropy calculations, operator $\mathcal{B}$ is introduced

$$\mathcal{B}[\Gamma] = -\Gamma \log_2(\Gamma)\,. \tag{2.15}$$

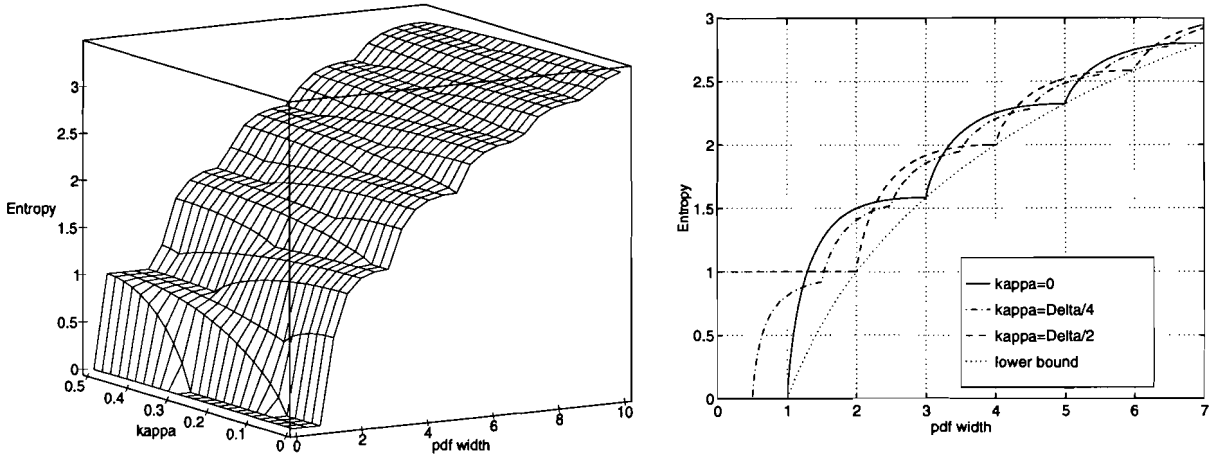
The entropy is the mean of the self-information

$$\begin{aligned} H(Q(x)) &= \sum_i -p_i I_i\,, \\ &= \sum_i \mathcal{B}[p_i]\,. \end{aligned} \tag{2.16}$$

For a uniform input distribution the following holds

$$\begin{aligned} H(Q(x)) &= \sum_i \mathcal{B}\left[\int_{(i-\frac{1}{2})\Delta+\kappa}^{(i+\frac{1}{2})\Delta+\kappa} \Pi_w(z)dz\right], \\ &= \begin{cases} 0, \text{for } i_L = i_R = 0, \text{ or otherwise;} \\ \mathcal{B}\left[\frac{(i_L+\frac{1}{2})\Delta+\kappa}{w} + \frac{1}{2}\right] + (i_R - i_L - 1)\mathcal{B}\left[\frac{\Delta}{w}\right] + \mathcal{B}\left[\frac{1}{2} - \frac{(i_R-\frac{1}{2})\Delta+\kappa}{w}\right], \end{cases} \end{aligned} \tag{2.17}$$

where i_L, i_R are given by (2.5). The entropy is plotted in Figure 2.3 for $\Delta = 1$, along with



(a) (b)

Figure 2.3: The entropy of a quantiser output, as a function of the pdf width and κ, for a uniform input pdf and Δ=1 (a). In (b) the entropy is plotted for $\kappa = 0, \frac{\Delta}{4}, \frac{\Delta}{2}$, in order to ease the comparison.

the lower bound for the entropy. This lower bound is achieved in the case that the input pdf is equally divided over N decision intervals, i.e.

$$\begin{aligned} H(Q(x)) &= \sum_i \mathcal{B}\left[\int_{(i-\frac{1}{2})\Delta+\kappa}^{(i+\frac{1}{2})\Delta+\kappa} \frac{1}{N\Delta} dz\right], \\ &= N\mathcal{B}\left[\frac{1}{N}\right], \\ &= \log_2 N\,. \end{aligned} \tag{2.18}$$

This lower bound can only be reached for $\kappa = 0$ or $\kappa = \frac{\Delta}{2}$, since these correspond with the mid-tread and mid-riser quantiser transfer functions, which are symmetric.

2.3 Undithered Rate-Distortion function

The rate is defined as the mean number of bits, necessary to represent a coded discrete source. The most important quantity is the rate R, necessary to guarantee a certain distortion D. Or, visa versa, the minimum distortion that can be achieved given a certain maximum rate. In general the rate can approximate the entropy as close as necessary, using complex encoding techniques. The rate is plotted as a function of the 2^{nd}-power distortion in Figure 2.4. In 2.4(d), R(D) is plotted is a function of Δ, while w is fixed and equals $\frac{1}{2\sqrt{3}}$, so that the input pdf has a variance of 1. It appears from Figure 2.4, that the rate-distortion function

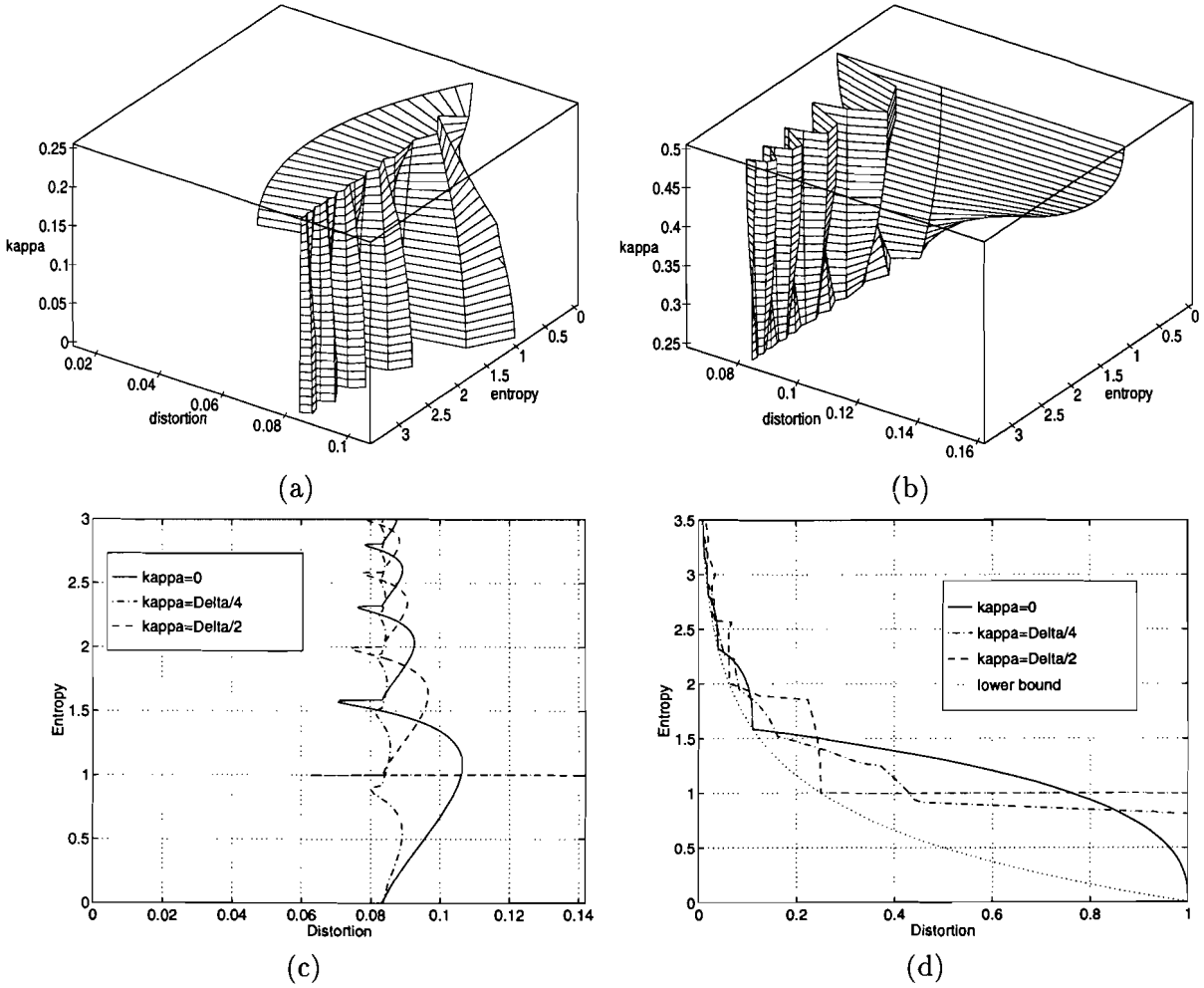


Figure 2.4: R(D) is plotted as a function of κ, for a uniform input pdf and Δ=1 in (a) and (b). In (c) R(D) is plotted for $\kappa = 0, \frac{\Delta}{4}, \frac{\Delta}{2}$, in order to ease the comparison. In (d) R(D) is plotted for a fixed w, and a variable quantising stepsize.

$R(D)$ has local optima. These optima appear in the case that the input pdf w equals even

and odd multiples of Δ for $\kappa = \frac{\Delta}{2}$ and $\kappa = 0$ respectively. If the statistical properties of the input signal are time-varying, the mentioned optima become critical. If $\kappa = \frac{\Delta}{4}$, $R(D)$ is relatively unsensitive for variations in the input pdf width w, given a certain constant Δ. The performance of an undithered quantiser with $\kappa = \frac{\Delta}{4}$ can not be judged however by comparing the 2^{nd} order rate-distortion functions, since also odd errors moments play a part, as stated in Section 2.1. The lower bound curve from Figure 2.4(d) is the $R(D)$ curve corresponding to the case that the input pdf is equally divided over N decision intervals. The distortion is then in agreement with the classical model (2.12), and the rate corresponds with (2.18). Substituting (2.12) in (2.18) yields the explicit relationship for $R(D)$

$$R(D) = \sqrt{\frac{w^2}{12D}}. \tag{2.19}$$

Clearly, this bound can only be reached with a mid-tread or a mid-riser quantiser, as stated in Section 2.2.

2.4 Entropy of the dithered signal

In general, the entropy of a quantised signal will increase if dither is applied. In this section the entropy analysis of the quantised RPDF input signal will be extended for the case that RPDF dither of 1 LSB amplitude is applied prior quantisation. The entropy of the dithered quantised signal can be calculated from (2.16) and (2.13) when p_x is replaced by $p_x \star p_\nu$

$$p_i = \int_{\alpha_{i-1}}^{\alpha_i} [p_x \star p_\nu](x)dx\,, \tag{2.20}$$

since adding two independent signals corresponds with the convolution of their pdf's.

The entropy of the dithered signal is plotted for $\kappa = 0, \frac{\Delta}{4}, \frac{\Delta}{2}$ in Figure 2.5 (a). It follows from this figure that the choice of κ does not affect the entropy of the dithered signal substantially for $w \geq \Delta$. For $w < \Delta$, the entropy gradually decreases to zero for $\kappa = 0$, whereas for $\kappa = \frac{\Delta}{4}, \frac{\Delta}{2}$ the entropy remains constant for small input pdf's. In the case of subtractive dithering, κ should always equal zero, since the choice of κ does not affect the distortion in any way.

The entropy of the dithered signal is plotted along with the entropy of the undithered signal for $\kappa = 0$ in Figure 2.5 (b). The difference between these two curves is referred to as the entropy loss, and is also depicted in Figure 2.5 (b). It follows from this figure that the entropy loss can be even more than 1 bit per sample for input pdf widths close to 1 LSB. This comparison is not quite fair however, as will be made clear by the following reasoning. Consider the case that the input pdf is rectangular and has a width not exceeding the quantiser stepsize Δ. If the quantiser used is of the mid-tread type, all input values will be truncated to zero in the undithered case. The performance appears to quite good however, judged by rate and distortion; the entropy equals zero and the distortion never exceeds $\frac{\Delta^2}{12}$. In the dithered case, the best performance is reached by subtractive dither, which results in a distortion identical to $\frac{\Delta^2}{12}$ and a rate up to 1.1 bits/sample. In practice, the subtractively dithered case provides

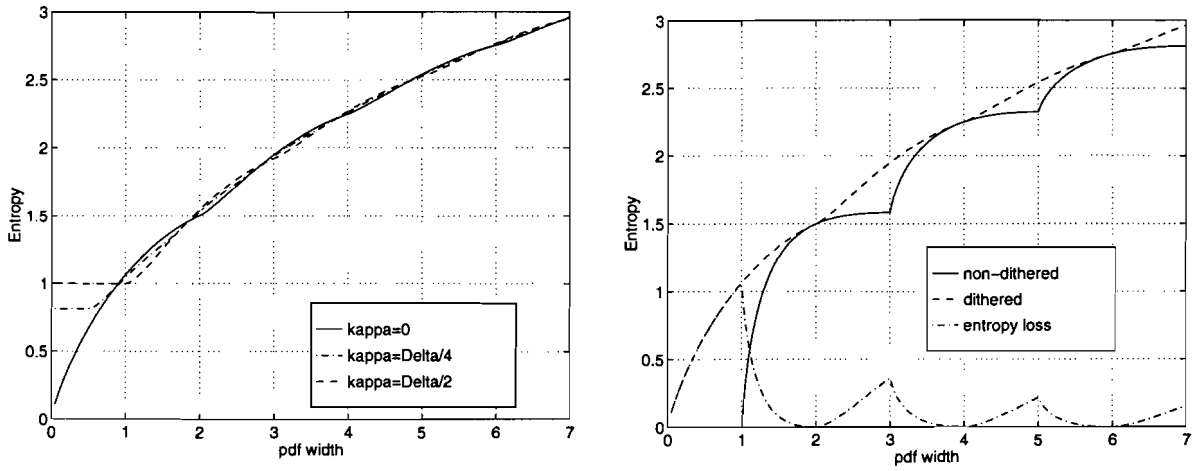


Figure 2.5: The entropy of the RPDF dithered signal is plotted for $\kappa = 0, \frac{\Delta}{4}, \frac{\Delta}{2}$ in (a). The entropy of the RPDF dithered and non-dithered signal, and the entropy loss, are plotted as a function of the pdf width in (b), with $\kappa = 0$ and Δ=1.

a quantised signal which consists of the original signal, free of artifacts, and a white noise component. The signal to noise ratio (SNR) approaches 1, as the input pdf approaches Δ, while the SNR equals zero in the undithered case. This example is most illustrative, but the same reasoning holds for coarse quantisation of a signal with an input pdf exceeding Δ. The undithered quantised signal then contains contours, which can be perceptually much more annoying than the white noise component in the subtractively dithered case.

Chapter 3

Reducing the entropy loss due to dithering

As discussed in Section 2.4, the entropy of the quantised signal increases excessively for coarse quantisation as a result of dithering. This is due to the fact that adding two independent signals corresponds with the convolution of their pdfs To clarify this, Figure 3.1 shows the input pdf convolved with the dither pdf resulting in a wider quantised pdf, which corresponds with a higher entropy. In this chapter an attempt is made to reduce the entropy by using different dither signals and a different coding strategy.

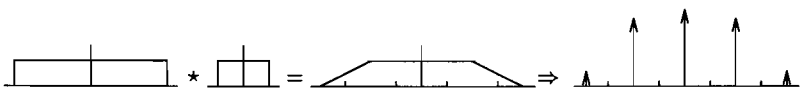


Figure 3.1: Pdf of the quantised dithered input signal

3.1 Alternative dither signals

A question of interest is if it is possible to reduce the amount of dither, so that the quantisation error modulation is partially eliminated at the cost of only a modest increase of the entropy. Obviously, such dithers do not obey the theorems in Appendix A. Fundamentally there are two ways of reducing the amount of dither; either the dithers amplitude is reduced, or the dither is only added occasionally. The first method will introduce noise intermodulation and is therefore not preferable. The latter method however, introduces a mixture of the properly dithered signal and the undithered signal.

3.2 The effect of applying the dither to the coding scheme

A preferable method would be the increase of the coding efficiency. In the entropy analysis of Section 2.4, the encoder and the decoder where assumed to have only $Q(x+\nu)$ to their disposal. But if, as in the case of subtractive dithering, the dither signal is known at both

the encoder and decoder side, the dither can be fed into the coding scheme as well. This will decrease the entropy since

$$H(Q(x+\nu)|\nu) \leq H(Q(x+\nu)) . \tag{3.1}$$

Consider the case that a signal with arbitrary pdf is quantised. If the statistics of the input signal are fully known or they are not varying in time, the entropy of the quantised signal equals

$$H(Q(x+\nu)|\nu) = \int_{-\infty}^{\infty} H(Q(x+\nu))p_\nu(\nu)d\nu . \tag{3.2}$$

If digital dither is used, $\nu = \{\nu_1, \nu_2, ..., \nu_N\}$, where the ν_i represent unique dither values. Thus equation (3.2) has the discrete equivalent

$$H(Q(x+\nu)|\nu) = \sum_{\nu_i \in \nu} H(Q(x+\nu_i))p_\nu(\nu_i) . \tag{3.3}$$

In general, an encoder can use the known statistics of the input signal, or it can determine these statistics itself by counting how often a certain input values occur. The first method will be indicated as fixed coding, and the latter will be indicated as adaptive coding.

3.2.1 Fixed coding

In the case of an stationary input signal with fully known statistical properties, the coder can calculate the statistics of $Q(x+\nu)$ given ν. This is done by shifting the original pdf p_x over ν and transforming the resulting pdf to an amplitude discrete pdf, to account for the quantising operation. Clearly, this calculation has to be performed each time instance. In the case that RPDF dither of 1 LSB top to top amplitude is applied, (3.2) can be simplified to

$$H(Q(x+\nu)|\nu) = \frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} H(Q(x+\nu))d\nu . \tag{3.4}$$

Note that (3.4) is independent of κ. This entropy is calculated for a uniform input distribution in Figure 3.2 (a). Also shown are the entropy of the dithered signal when the dither is not applied to the coder, the entropy of the undithered mid-tread quantised signal, and the entropy loss due to dithering for both cases; the coder is or is not provided with the dither. It follows from Figure 3.2 (a) that a gain in entropy up to 0.37 bits per sample can be achieved by passing the dither to the coding scheme. This gain is achieved for an uniform input pdf of 1 LSB amplitude, for which the entropy decreases from 1.06 to 0.69 bit. Note that, compared to the undithered mid-tread quantised case, the entropy can be locally smaller than the undithered entropy.

3.2.2 Adaptive coding

It is clear that fixed coding results in the best possible coding efficiency, but in general this method is not robust since the statistical properties of the input signal are mostly not exactly known or they vary in time. An adaptive coder keeps track of the statistics by simply counting how often the distinct input values occur. When the dither is applied to the coder, statistics of $Q(x+\nu)$ must be determined given ν. Clearly, the convergence rate then decreases with

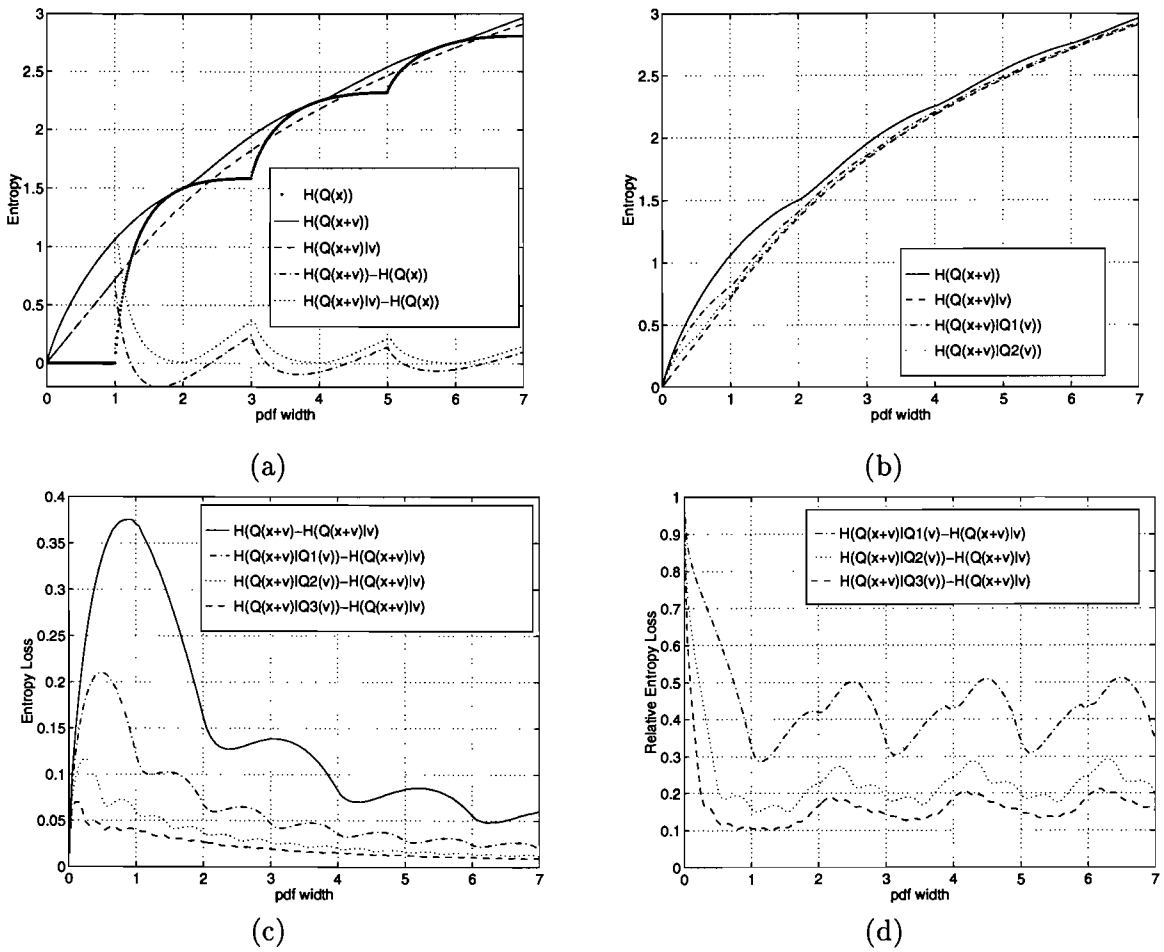


Figure 3.2: Entropy if the dither is applied to the encoder (a), Entropy if the dither is quantised in 2 and 4 levels respectively, before being applied to the encoder (b), Entropy loss due to quantisation of the dither (c), Entropy loss due to quantisation of the dither, relative to the entropy loss if no dither was applied to the coder (d).

the number of discrete dither values N, while the coder complexity increases with N. In general N is large, say 2^8, which causes poor tracking capabilities. A solution to this problem is to pass a coarsely quantised version of the dither to the coder at the cost of a lower coding efficiency. The corresponding coding scheme is depicted in Figure 3.3. The entropy then equals

$$H(Q(x+\nu)|Q'(\nu)) = \sum_{\nu_j \in Q'(\nu)} H(Q(x+\nu)|Q'(\nu)=\nu_j) p_{Q'(\nu)}(\nu_j)\,, \tag{3.5}$$

where the ν_j represent all possible distinct quantised dither values. To clarify this, the quantised pdf's given sign(ν) are depicted in Figure 3.4. The entropy then follows from (3.5)

$$H((Q(x+\nu)|\text{sign}(\nu)) = H(Q(x+\nu)|\nu \geq 0) p_\nu(\nu \geq 0) + H(Q(x+\nu)|\nu < 0) p_\nu(\nu < 0)\,. \tag{3.6}$$

The entropy is calculated for a uniform input pdf for the cases that no dither, 1 and 2 bit quantised versions of the dither, and the continuous dither are applied to the coder respect-

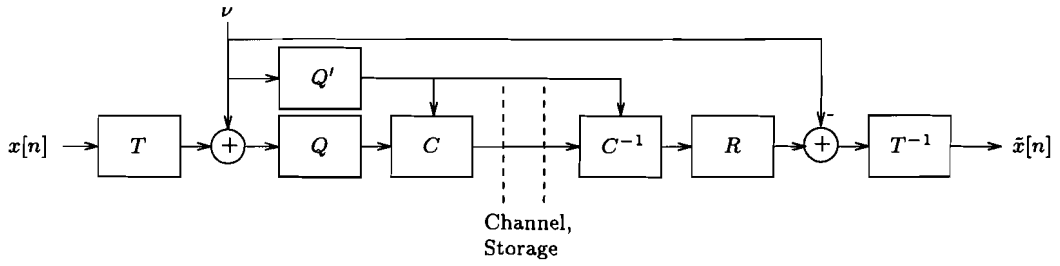


Figure 3.3: Blocks of a dithered source coding system

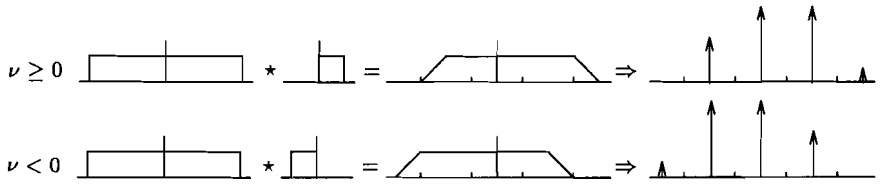


Figure 3.4: Quantised pdf, the encoder is provided with sign(ν)

ively. The result is plotted in Figure 3.2 (b). The 1 and 2 bit quantisation are indicated as $Q1$ and $Q2$. It follows from this figure that the entropy converges to the lower bound rather fast. In Figure 3.2 (c), the loss in entropy of passing no dither, 1, 2, or 3 bit quantised dither to the coder instead of the continuous dither. Figure 3.2 (d) shows the entropy loss relative to the entropy loss if no dither was applied to the coder. It follows from this plot, that passing only sign(ν) to the coder results in about 60% of the achievable entropy gain. Passing the four level quantised dither to the coder results in 80% of the achievable gain. Quantising the dither less coarse, for example in eight levels, does not increase this gain much more. The optimal number of levels is determined by the input signal; if its the statistical properties slowly vary in time, a high number of levels is preferable and visa versa.

Chapter 4

Rate-Distortion curve for Generalised Gaussian sources

As mentioned in Chapter 2, an audio signal can roughly be modelled by a Laplacian distribution. A better estimate, however, is given by a Generalised Gaussian distribution [11]. The distribution of the samples a subband decomposition scheme can also be modeled with a Generalised Gaussian pdf (GG-pdf). Except for the DC-term, the frequency coefficients generated by a transform coding technique (e.g. Discrete Cosine Transform), are GG-distributed. Further investigation shows that the difference between adjacent DC-values can be also modelled with a GG-pdf. Also differential image histograms can be modelled with a GG-pdf [13].

For a subset of the Generalised Gaussian probability density functions, the rate and distortion can be calculated analytically. This is shown for a Laplacian distribution in Section 4.2. In the previous chapter, the entropy loss due to dithering was investigated for uniformly distributed signals, and this loss was reduced by applying the dither to the coding scheme. This investigation will be extended to Generalised Gaussian distributed signals in this chapter.

4.1 The Generalised Gaussian probability density function

The GG-pdf is given by

$$p_x(x;\mu,\sigma^2,\gamma) = ae^{-[b|x-\mu|]^\gamma}\,,\ x\in\mathbb{R}\,, \tag{4.1}$$

where, μ,σ^2,γ are mean, variance and shape parameter of the distribution respectively. The positive constants a and b are given by

$$\begin{aligned} a &= \frac{b\gamma}{2\Gamma(1/\gamma)}\,,\\ b &= \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}\,, \end{aligned} \tag{4.2}$$

where $\Gamma(x)$ is the Gamma function given by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \tag{4.3}$$

The Generalised Gaussian distribution is plotted in Figure 4.1 for γ=0.5, γ=1 (Laplacian pdf), and γ=2 (Gaussian pdf), where 'shape' in the legend corresponds with γ. All three

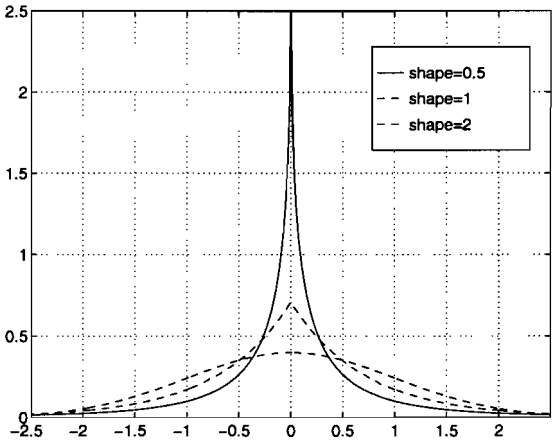


Figure 4.1: Generalised Gaussian pdf for $\gamma = 0.5, \gamma = 1$ and $\gamma = 2$.

distributions shown have a variance of 1. A closed form expression for both rate and distortion can be obtained for uniform quantisation in the case that $\frac{1}{\gamma}$ is a positive integer. This is because in both calculations the probability density function needs to be integrated. It can easily be seen that this calculation can only be performed analytically for $n = 0, 1, 2, ...$

$$\int e^{-x^\gamma} dx = \tfrac{1}{\gamma}(-1)^{\frac{1}{\gamma}} \int e^Q Q^n dQ\,, \tag{4.4}$$

$$= (n+1)(-1)^{n+1} e^Q \sum_{j=0}^{n} \frac{n!}{(n-j)!}(-1)^j Q^{n-j}\,, \tag{4.5}$$

where $Q = -x^\gamma$, and $n = \frac{1}{\gamma} - 1$. As a result, $\frac{1}{\gamma}$ must be a positive integer. In practice, the complexity of closed-form expressions for rate and distortion increases rapidly as $\frac{1}{\gamma}$ is greater than 1.

4.2 Rate and Distortion for a Laplacian source

In this section, a closed-form expression is derived for both rate and distortion of a uniformly quantised memoryless Laplacian source. The Laplacian probability density function is found by substituting $\gamma = 1$ in the generalised Gaussian distribution

$$P_x(x) = \frac{1}{\sigma_x\sqrt{2}} e^{-\sqrt{2}\frac{|x|}{\sigma_x}}\,. \tag{4.6}$$

4.2.1 Undithered Distortion and Entropy

The distortion is obtained by substituting (4.6) in (2.3)

$$D = \sigma_x^2 - \alpha e^{-\frac{\Delta}{\sigma_x\sqrt{2}}} + \left(\beta e^{\frac{\Delta}{\sigma_x\sqrt{2}}} - \alpha e^{-\frac{\Delta}{\sigma_x\sqrt{2}}}\right) \frac{e^{-\frac{\sqrt{2}\Delta}{\sigma_x}}}{1 - e^{-\frac{\sqrt{2}\Delta}{\sigma_x}}}\,, \tag{4.7}$$

where

$$\alpha = \frac{\Delta^2}{4} + \frac{\Delta\sigma_x}{\sqrt{2}} + \sigma_x^2,$$
$$\beta = \frac{\Delta^2}{4} - \frac{\Delta\sigma_x}{\sqrt{2}} + \sigma_x^2. \tag{4.8}$$

The entropy is defined by (2.16), where the p_i are calculated from (2.13)

$$p_i = \begin{cases} 1 - e^{-\frac{\Delta}{\sigma_x\sqrt{2}}}, & \text{for i=0} \\ e^{-\frac{\sqrt{2}|i|\Delta}{\sigma_x}} \sinh(\frac{\Delta}{\sigma_x\sqrt{2}}), & \text{otherwise} \end{cases}. \tag{4.9}$$

Using

$$\sum_{n=-\infty}^{\infty} a^{-n} = \frac{a}{1-a},$$
$$\sum_{n=-\infty}^{\infty} n a^{-n} = \frac{a}{(1-a)^2}, \tag{4.10}$$

and using the symmetry of (4.9) yields

$$\begin{aligned} H(Q(x)) &= \mathcal{B}[p_0] + 2\sum_{i=1}^{\infty} \mathcal{B}[p_i], \\ &= \mathcal{B}[p_0] + 2\left[\frac{A\mathcal{B}[C]}{1-A} + \frac{C\mathcal{B}[A]}{(1-A)^2}\right], \end{aligned} \tag{4.11}$$

where

$$A = e^{-\frac{\Delta\sqrt{2}}{\sigma_x}}, \tag{4.12}$$
$$C = \sinh\left(\frac{\Delta}{\sigma_x\sqrt{2}}\right). \tag{4.13}$$

4.2.2 Dithered Distortion and Entropy

In this section the influence of subtractive dithering is considered with respect to the distortion and the entropy. Dithering with RPDF noise and quantising are accounted for by convoluting the input pdf with Π_Δ and $\Delta\Pi_\Delta$

$$p_i = \Delta[p_x \star \Pi_\Delta \star \Pi_\Delta](i\Delta), \tag{4.14}$$

where $p_x(x)$ is given by (4.6). The p_i are given by

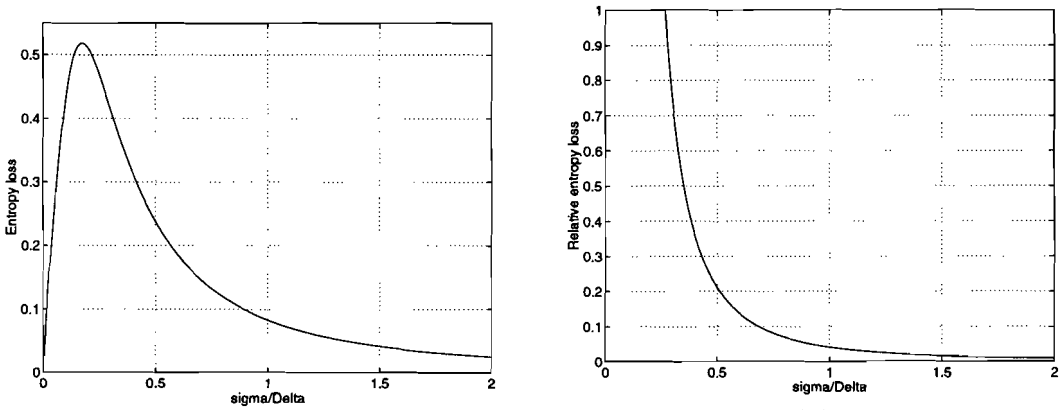
$$p_i = \begin{cases} 1 - \frac{\sigma_x}{\sqrt{2}\Delta}\left(1 - e^{\frac{\Delta\sqrt{2}}{\sigma_x}}\right), & \text{for i=0} \\ \frac{\sigma_x}{\Delta\sqrt{2}}\left(\cosh\left(\frac{\Delta\sqrt{2}}{\sigma_x}\right) - 1\right) e^{-\frac{\sqrt{2}|i|\Delta}{\sigma_x}}, & \text{otherwise} \end{cases}. \tag{4.15}$$

The entropy then follows from substituting (4.15) in (4.11)

$$H(Q(x)) = \mathcal{B}[p_0] + 2\left[\frac{A\mathcal{B}[E]}{1-A} + \frac{E\mathcal{B}[A]}{(1-A)^2}\right], \tag{4.16}$$

where A is given by (4.12), and E is given by

$$E = \frac{\sigma_x}{\Delta\sqrt{2}} \left(\cosh\left(\frac{\Delta\sqrt{2}}{\sigma_x} \right) - 1 \right) . \tag{4.17}$$



(a) (b)

Figure 4.2: Entropy loss due to RPDF dithering, Laplacian source (a), Entropy loss due to RPDF dithering, relative to the entropy if no dither was applied (b).

The increase in entropy due to dithering is plotted in Figure 4.2.2 for a Laplacian pdf. Note that the entropy is fully determined by the ratio $\frac{\sigma}{\Delta}$, e.g. magnifying the input signal and the quantiser step size with the same factor results in the same bit rate. Figure 4.2.2a shows that entropy losses of more than 0.5 bit per sample can occur. Furthermore, the relative increase in entropy is plotted in Figure 4.2.2(b) and indicates that the entropy increases more than 100% for signals with a standard deviation of less than 30% of the quantiser stepsize.

4.3 Deadzone Quantised Dithered R(D)

It was shown in the previous section that dithering a uniform quantiser causes the entropy to increase excessively for signals with a small standard deviation with respect to the quantiser stepsize. In order to avoid this, the use of a deadzone quantiser $Q_\alpha(.)$ is considered in this section. The transfer function of this quantiser is depicted in Figure 4.3.

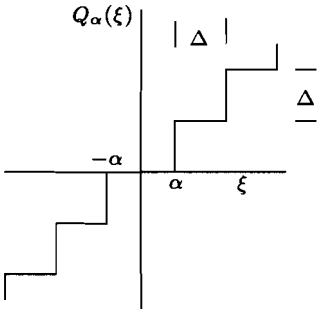


Figure 4.3: Deadzone-quantiser characteristic

The deadzone-width equals 2α. In the following, $\alpha \geq \frac{\Delta}{2}$.

4.3.1 Deadzone Quantised Dithered Distortion

The output of a subtractively dithered quantiser equals $q_\alpha(x,\nu) = Q_\alpha(x+\nu) - \nu$. The distortion of a RPDF dithered deadzone quantiser is defined by

$$\begin{aligned} E[(\tilde{x}-x)^2] &= \int_{-\infty}^{\infty} p_x(x)p_\nu(\nu)(q_\alpha(x,\nu)-x)^2 d\nu dx\,, \\ &= \int_{-\infty}^{\infty} p_x(x)\left[\tfrac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}(q_\alpha(x,\nu)-x)^2 d\nu\right] dx\,. \end{aligned} \tag{4.18}$$

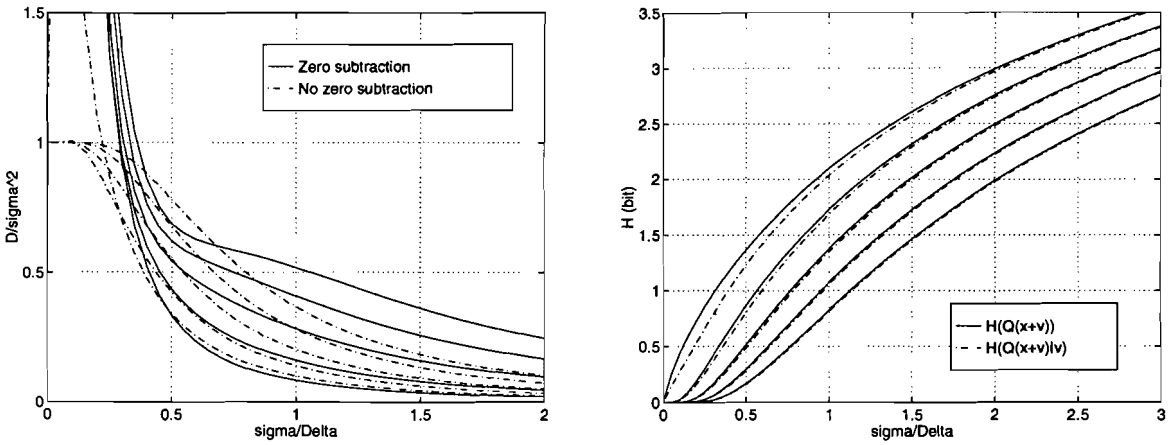
Now the term between brackets in (4.18) is evaluated:

$$\begin{aligned} |x| < \alpha - \tfrac{\Delta}{2} \quad &: \quad [.] = x^2 + \frac{\Delta^2}{12}\,, \\ \alpha - \tfrac{\Delta}{2} \leq |x| \leq \alpha + \tfrac{\Delta}{2} \quad &: \quad [.] = \frac{\Delta^2}{12} + (\tfrac{\Delta}{2} - \alpha)(x^2 - |x|(\tfrac{1}{2} + \tfrac{\alpha}{\Delta})) = f(x)\,, \\ |x| > \alpha + \frac{\Delta}{2} \quad &: \quad [.] = \frac{\Delta^2}{12}\,. \end{aligned} \tag{4.19}$$

Using that the pdf is symmetrical, i.e. $p_x(x) = p_x(-x)$, (4.18) only needs to be evaluated for positive x:

$$\begin{aligned} E[(\tilde{x}-x)^2] &= 2\int_0^{\alpha-\frac{\Delta}{2}}(x^2 + \tfrac{\Delta^2}{12})p_x(x)dx + 2\int_{\frac{\Delta}{2}-\alpha}^{\frac{\Delta}{2}+\alpha} f(x)p_x(x)dx + 2\int_{\frac{\Delta}{2}+\alpha}^{\infty}\frac{\Delta^2}{12}p_x(x)dx\,, \\ &= 2\int_0^{\frac{\Delta}{2}-\alpha} x^2 p_x(x)dx + 2\int_{\frac{\Delta}{2}-\alpha}^{\frac{\Delta}{2}+\alpha}(f(x) - \tfrac{\Delta^2}{12})p_x(x)dx + \frac{\Delta^2}{12}\,. \end{aligned} \tag{4.20}$$

As a result, the integral is finite, so that it is numerical traceable. The distortion is calculated for a deadzone of $\{\Delta, \frac{3}{2}\Delta, \ldots \frac{5}{2}\Delta, 3\Delta\}$, and is plotted in Figure 4.4(a), indicated with 'zero subtraction'. This figure shows that the distortion becomes even larger than the signal power σ^2 for $\Delta > 3\sigma$. To overcome this problem, it can be decided not to subtract the dither from the quantiser output in the case that the quantiser output equals zero. Signals with a small



(a) (b)

Figure 4.4: Rate and Distortion of a Laplacian source for several deadzone widths

standard deviation with respect to the quantiser stepsize are quantised to zero. As a result, the resulting noise power will approximately equal the signal power in stead of $\frac{\Delta^2}{12}$. Equation 4.18 now yields

$$\begin{aligned} |x| < \alpha - \tfrac{\Delta}{2} \quad &: \quad [.] = x^2 \,, \\ \alpha - \tfrac{\Delta}{2} \leq |x| \leq \alpha + \tfrac{\Delta}{2} \quad &: \quad [.] = -\tfrac{2}{3\Delta}x^3 + \tfrac{1}{2}x^2 + \tfrac{\alpha^2}{\Delta}x + \tfrac{\Delta^2}{24} - \tfrac{\alpha^3}{3\Delta} = g(x) \,, \\ |x| > \alpha + \tfrac{\Delta}{2} \quad &: \quad [.] = \frac{\Delta^2}{12} \,. \end{aligned} \tag{4.21}$$

Using that the pdf is symmetrical, i.e. $p_x(x) = p_x(-x)$, (4.18) only needs to be evaluated for positive x

$$\begin{aligned} E[(\tilde{x} - x)^2] &= 2\int_0^{\frac{\Delta}{2}-\alpha} x^2 p_x(x)dx + 2\int_{\frac{\Delta}{2}-\alpha}^{\frac{\Delta}{2}+\alpha} g(x)p_x(x)dx + 2\int_{\frac{\Delta}{2}+\alpha}^{\infty} \frac{\Delta^2}{12}p_x(x)dx \,, \\ &= \frac{\Delta^2}{12} + 2\int_0^{\frac{\Delta}{2}-\alpha} (x^2 - \tfrac{\Delta^2}{12})p_x(x)dx + 2\int_{\frac{\Delta}{2}-\alpha}^{\frac{\Delta}{2}+\alpha} (g(x) - \tfrac{\Delta^2}{12})p_x(x)dx \,. \end{aligned} \tag{4.22}$$

This integral is also numerical traceable. The distortion is calculated for a deadzone of $\{\Delta, \frac{3}{2}\Delta, \ldots \frac{5}{2}\Delta, 3\Delta\}$, and is plotted in Figure 4.4(a), indicated with 'no zero subtraction'. It follows from this figure that the distortion never exceeds σ^2 if a deadzone of $\frac{3}{2}\Delta$ or more is used and the dither is not subtracted for a zero quantiser output. Note that the bit rate is not affected using 'no zero subtraction'.

4.3.2 Deadzone Quantised Dithered Entropy

The calculation of the entropy of a signal quantised with a dithered deadzone quantiser can be performed using (4.14). This can be seen from Figure 4.5, where the ticks represent the decision intervals. The middle bar corresponds with the deadzone quantiser. The p_i along this bar can also be derived from the upper and lower bars, which represent uniform quantisers,

shifted to the right and to the left respectively. Only p_0 cannot be obtained in this way, but since the sum of all p_i equals 1, p_0 is also known. Note that shifting a uniform quantiser over a multiple times Δ yields the same p_i, only the p_i need to be reindexed. As a consequence, only shifts of $\{-\frac{\Delta}{2} \ldots \frac{\Delta}{2}\}$ have to be performed and the p_i are stored in a lookup table (lut). The entropies of all deadzone quantisers can be calculated from this lut. The resulting entropy is plotted in Figure 4.4(b) with solid lines.

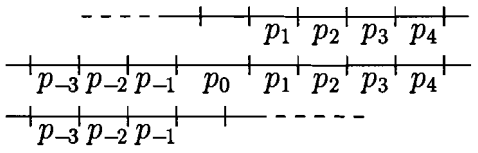


Figure 4.5: Calculation of deadzone-quantiser entropy

In Chapter 3 was shown that the entropy can be reduced by applying the dither to the coder. The entropy can then be calculated from (3.4). In order to reduce the number of calculations, this equation can be approximated with

$$H(Q(x+\nu)|\nu) \approx \tfrac{1}{N} \sum_{i=1}^{N} H(Q(x+\nu_i)), \tag{4.23}$$

where the ν_i are equally spaced over the interval $\{-\frac{\Delta}{2} \ldots \frac{\Delta}{2}\}$. For large N, this approximation is justified. The entropy can now again be calculated from the lut mentioned, using reindexing and shifting over S

$$p_i = \Delta[p_x \star \Pi_\Delta \star \Pi_\Delta](i\Delta + S). \tag{4.24}$$

The shifts needed for the negative and positive p_i's are given by S_{neg} and S_{pos} in

$$\begin{aligned} S_{neg} &= -\tfrac{\nu}{\Delta} + \alpha - \tfrac{\Delta}{2}, \\ S_{pos} &= -\tfrac{\nu}{\Delta} - \alpha + \tfrac{\Delta}{2}. \end{aligned} \tag{4.25}$$

Normalising these shifts to the interval $\{-\frac{\Delta}{2} \ldots \frac{\Delta}{2}\}$ yields the shifts T_{neg} and T_{pos}

$$\begin{aligned} T_{neg} &= S_{neg} - \Delta \left\lfloor \tfrac{S_{neg}}{\Delta} + \tfrac{1}{2} \right\rfloor, \\ T_{pos} &= S_{pos} - \Delta \left\lfloor \tfrac{S_{pos}}{\Delta} + \tfrac{1}{2} \right\rfloor. \end{aligned} \tag{4.26}$$

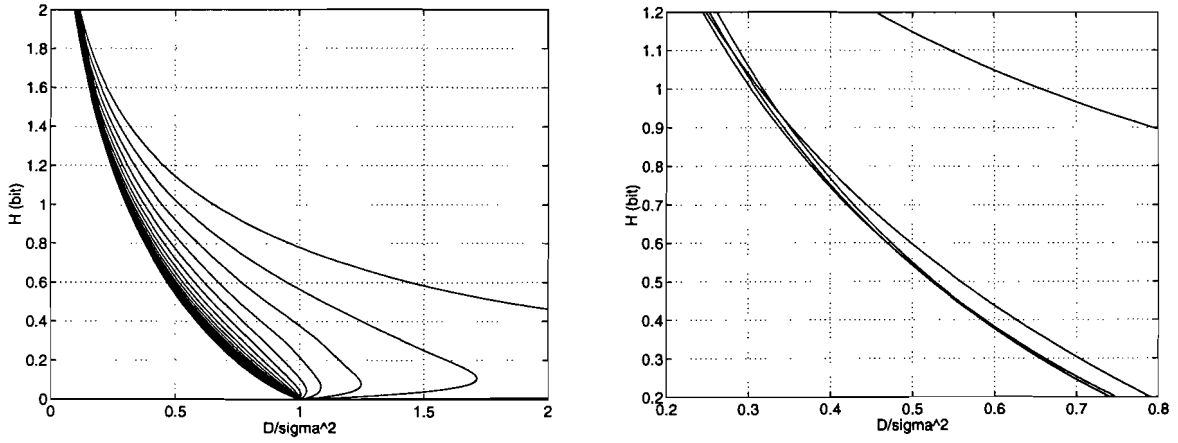
Now the remaining shift is achieved by adding N_{neg} and N_{pos} to the indices i.

$$\begin{aligned} N_{neg} &= \tfrac{1}{\Delta}(T_{neg} - S_{neg}), \\ N_{pos} &= \tfrac{1}{\Delta}(T_{pos} - S_{pos}). \end{aligned} \tag{4.27}$$

The entropy $H(Q(x+\nu)|\nu)$ can thus also be calculated from the lut, and is also plotted in Figure 4.4(b) with dashed lines. From this figure it follows that gain in entropy is small if the deadzone width is $\frac{3}{2}\Delta$ or more. For this reason, $H(Q(x+\nu))$ will be used in the sequel.

4.3.3 Deadzone Quantised Dithered R(D)

Combining the 'no zero subtraction' distortion with the entropy $H(Q(x+\nu))$ for $\gamma = 1$ of Figure 4.4 yields the rate-distortion curves in Figure 4.6. In (a), R(D) curves are plotted for 41 deadzone widths in between Δ and 3Δ. In (b), 5 of these curves are magnified. The



(a) (b)

Figure 4.6: R(D) of a Laplacian source for 41 deadzones in $\{\Delta \ldots 3\Delta\}$ (a) and R(D) of a Laplacian source for 5 deadzones in $\{\Delta \ldots 3\Delta\}$ magnified (b)

magnification shows that the curves cross each other. Thus, the deadzone that yields a minimal distortion is a function of the wanted bit rate H. This function is examened for a Laplacian source ($\gamma = 1$), but also for sources with $\gamma = \frac{1}{2}$ and $\gamma = \frac{3}{2}$. The results are plotted in Figure 4.7. In the legend of this plot, 'shape' corresponds with γ. Note that the choice

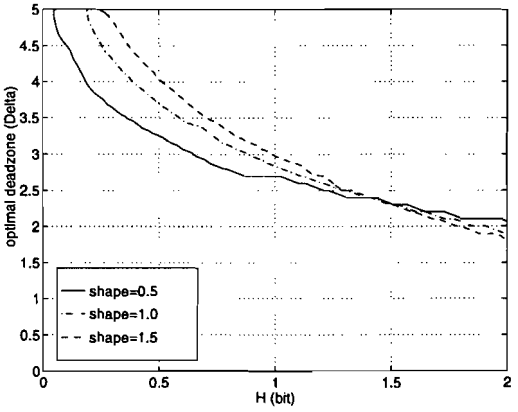


Figure 4.7: Optimal deadzone as a function of the entropy for $\gamma = \frac{1}{2}, 1, \frac{3}{2}$

of the deadzone width is not critical, since the all R(D) curves with a deadzone of 2Δ or higher are very close to each other in Figure 4.6. The distribution of the AC coefficients and differential DC coefficients of the DCT transformed 'Lenna' image that will be used in Chapter 9 is fitted with a GG-pdf in Figure 4.8. The AC coefficients are distributed with γ=0.28 and

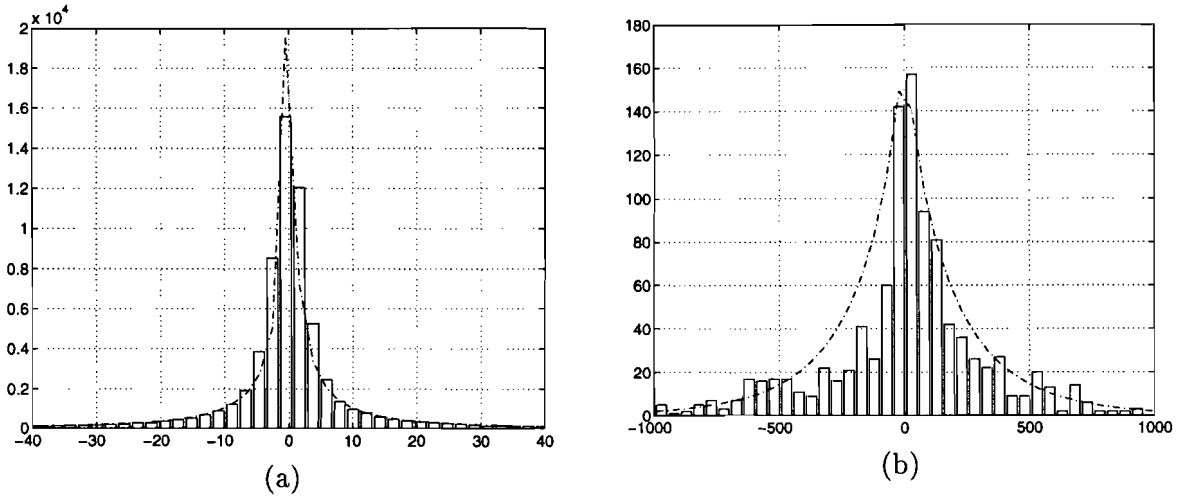


Figure 4.8: Distribution of the 'Lenna' DCT AC and differential DC coefficients

the differential DC coefficients are distributed with γ=0.84. Note that the AC coefficients resemble the GG-distribution more closely than the DC coefficients. This is because the DC-distribution is calculated from 1024 coefficients only, while the AC-distribution is calculated from 64512 coefficients. The optimal deadzone can be estimated by extrapolating Figure 4.7. The optimal deadzone width approximately equals 3Δ for $\gamma = 0.28$ and a desired bitrate of 0.5 bbp.

Chapter 5

Noise shaping and shaped dither

5.1 Noise shaping

Noise shaping is a technique which casts the quantisation noise into less perceptible frequency regions. The quantisation error is filtered by the noise shaping filter $H(z)$, and subtracted from the quantiser input. The quantisation noise is then shaped according to

$$\tilde{X} = X + E(1 - H(z)) . \tag{5.1}$$

In the design of the noise shaping filter, the quantisation error is interpreted as a uniform distributed white noise signal, as discussed in Section 1.2. The assumption that the noise spectrum is flat is not justified for coarse quantisation. As a result, the desired spectral shape of the quantisation noise is no longer guaranteed by properly choosing $H(z)$. Furthermore, the error remains a function of the input signal x. Due to this dependence, the output of the noise shaping scheme consists of a constantly repeated sequence when a constant value is applied to the input. These perceptually annoying sequences are called limit cycles.

5.2 Dithered noise shaping

Dither, either subtractive or non-subtractive, breaks up these limit cycles and gives the total error a random nature. Subtractive dither, renders the quantisation error uniform distributed and independent of the input signal. Thus, a subtractively dithered quantiser can be modelled by the classical model. As a result, the shaped error spectrum will exactly resemble the transfer function of the noise shaping scheme, $1 - H(z)$. A subtractively dithered noise shaping scheme is depicted in Figure 5.1. It can be proven that the noise shaping scheme of Figure 5.1 is equivalent to that of Figure 5.2. This is an important result, since it follows from Figure 5.2 that existing implementations of quantisers equipped with a noise shaping filter can be used combined with subtractive dither. The noise shaping scheme produces an overall error ϵ which is spectrally shaped with respect to the quantisation error e according to

$$PSD_\epsilon(f) = |1 - H(e^{-j2\pi fT})|^2 PSD_e(f) , \tag{5.2}$$

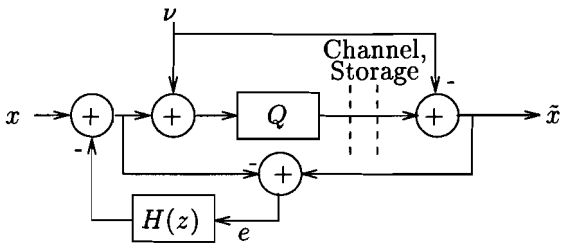


Figure 5.1: Subtractively dithered noise shaping scheme

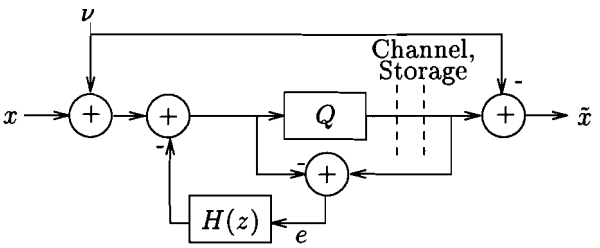


Figure 5.2: Equivalent subtractively dithered noise shaping scheme

where $H(e^{-j2\pi fT})$ represents the frequency response of the noise shaping filter $H(z)$. The power spectral density of the quantising error in the unshaped scheme is

$$PSD_e(f) = \frac{\Delta^2 T}{6}, \tag{5.3}$$

for the subtractively dithered case, and

$$PSD_e = PSD_\nu(f) + \frac{\Delta^2 T}{6}, \tag{5.4}$$

in the non-subtractively dithered case. There are no possible equivalent schemes where non-subtractive dither is applied outside the noise shaping scheme.

5.3 Shaped dither

Another way of shaping the quantisation error, is by shaping the dither. In most publications, the dither is assumed to be independent of the input signal. Lipshitz et al. investigated the existence of a non-subtractive dither which is not white [8]. One such dither is the so called 'high-pass' TPDF dither, which is generated by shaping white RPDF dither with a FIR filter with coefficients {1,-1}. Using such non-white TPDF dither has the advantage over white TPDF dither that only one random noise generator is needed instead of two, and that the dither is shaped to the higher and less audible frequencies. Furthermore the conditional entropy decreases as a consequence of the correlated dither, compared with uncorrelated TPDF

dither. In a quantising system without feedback and using a shaped non-subtractive RPDF dither, the coefficients of the dither shaping FIR filter c_i must be chosen such that the first or the last non-zero coefficient is a non-zero integer, and there are at least two distinct non-zero integer coefficients in total.

For example:

- highpass dither {...,0,0,0,1,-1,0,0,0,...},
- notch dither {...,0,0,0,1,-1,1,-1,1,0,0,0,...},
- but also {...,0,0,0,1,$-\frac{1}{2}$,1,$-\frac{1}{2}$,0,0,0,...},

all satisfy the necessary conditions. The power spectra of the highpass dither and the notch dither are plotted in Figure 5.3. The highpass dither is shaped to the higher frequencies, while the notch dither suppresses the 4kHz and 13kHz regions. The 13kHz region is especially important for headphone use. The highpass and notch dither will be used in the audio experiments.

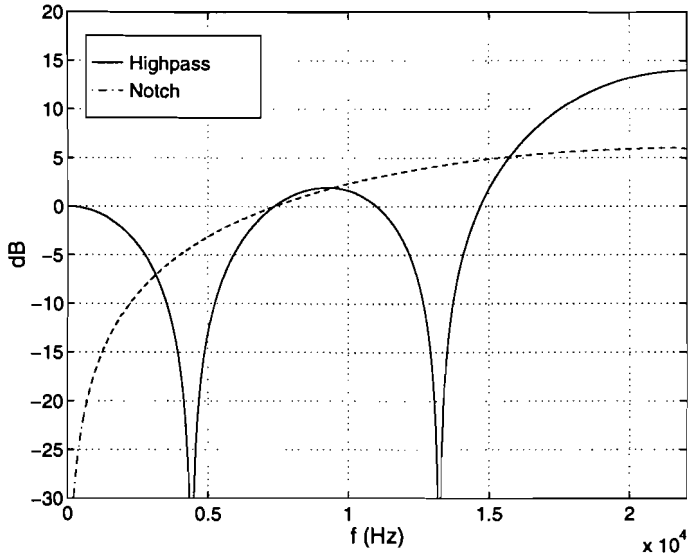


Figure 5.3: Shaped dither characteristics

S

Chapter 6

Subband Coding

Subband Coding (SBC) is a widely used technique in audio coding. In this chapter a four band SBC system will be discussed. Dither and noise shaping are incorporated in the quantisation of the subband signals.

6.1 Dither and Subband Coding

A general dithered subband coding scheme is shown in Figure 6.1.

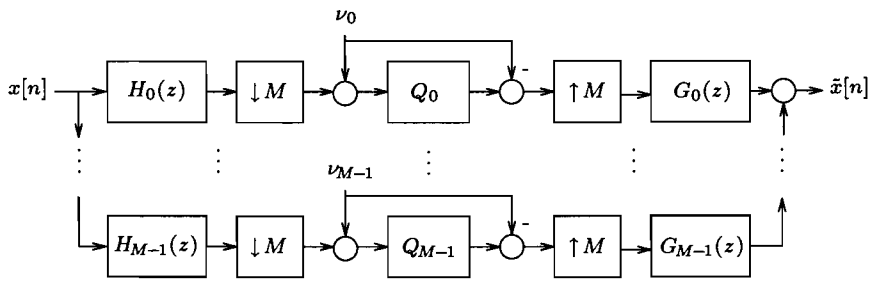


Figure 6.1: Implementation dithered Subband Coding scheme

In the analysis filter bank, the bandpass filters $\{H_0 \ldots H_{M-1}\}$ decompose the input signal into M frequency subbands. Typically, these filters are FIR filters of length L. Steep bandpass filters can be designed if L is chosen to be large with respect to M. A disadvantage, however, is that pre-echos can appear in audio applications. In video applications, these artifacts appear as ringing. The ratio $\frac{L}{M}$ is typically chosen around 16 for audio applications. The output of each filter is critically down-sampled. After down-sampling, the subband signals are quantised using subtractive dither. In the synthesis bank, the subband signals are up-sampled, inverse filtered, and summed, to obtain the reconstructed signal $\tilde{x}[n]$. The bandpass filters that are used for the experiments are depicted in Figure 6.2. The number of bands (M) equals 4 and the FIR filters are of length $L = 4 \cdot 16$. The separation into subbands allows for efficient coding. In addition, the decomposition of the signal into several subbands allows for the usage of different bit rates and coding methods. In this way even better coding

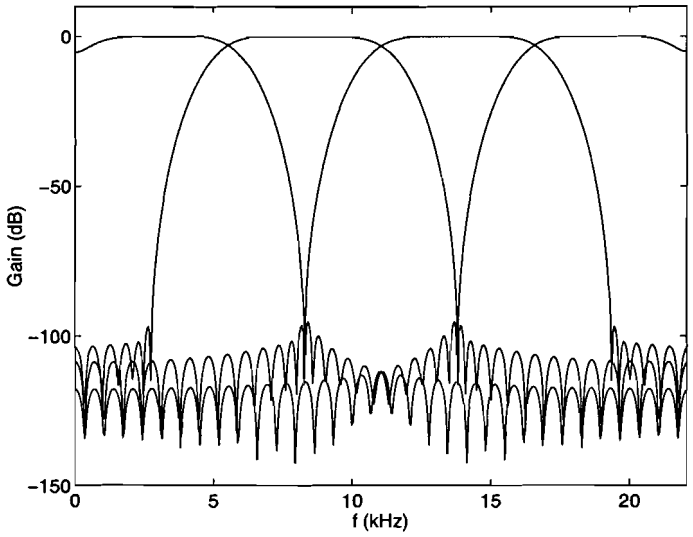


Figure 6.2: Subband filters for $M = 4$

performances can be achieved. Improvements using dither combined with subband coding where reported by Chen [10] for video coding. The two lowest subbands in a four subband system where dithered.

6.2 Subband Coding and Noise Shaping

The noise shaping technique discussed in Chapter 5, can also be applied to subband signals. The shaping should be performed in such a way, that the perceived noise power at the output of the filter bank is minimal. In the following, noise will be shaped according to the hearing threshold in quiet. This hearing threshold is plotted in Figure 6.3. When the noise power exceeds this threshold, it becomes audible for a silent input signal. More sophisticated models of psychoacoustical weighting also make use of the masking properties of the audio signal. This is not discussed in this report. If dither is used in the subbands, the flat noise floor can be shaped according to Figure 6.3. 9 points of the hearing curve where taken per subband, and a fifth-order noise shaping FIR filter was designed under MSE constraint. The resulting noise shaping spectra are also shown in Figure 6.3. Since the hearing curve is clipped at 60dB, no noise shaping curve was designed for the fourth subband. The filter coefficients of the noise shaping filters used are:

$$
\begin{aligned}
FIR1 = &\ \{-0.305259 \quad -0.581887 \quad -0.089763 \quad -0.186731 \quad -0.011653\}, \\
FIR2 = &\ \{\ 0.745519 \quad -0.420158 \quad\ \ 0.324080 \quad -0.173283 \quad\ \ 0.127672\}, \\
FIR3 = &\ \{\ 2.506883 \quad -3.041073 \quad\ \ 2.391820 \quad -1.222848 \quad\ \ 0.325896\}, \\
FIR4: &\ \text{no noise shaping.}
\end{aligned}
$$

In fact, the error is shaped with $1 - H(z)$ as in (5.1), where $H(z)$ is the transfer function of the noise shaping filters. These filters are minimum phase filters, so that the stability of the

synthesis filters is guaranteed. The filters originally have an average gain of

$$\int \log_{10} |1 - H(e^{j\theta})|^2 d\theta = 0\text{dB}\,, \tag{6.1}$$

where the integral is taken over the subband of interest. The noise power in band 1, 2 and 3 are chosen to be 6.4dB, 38dB, 60dB higher respectively than the noise power in band 0, so that the noise shaping spectra match with the hearing curve. The filter functions plotted in 6.3 are already shifted by these gain factors.

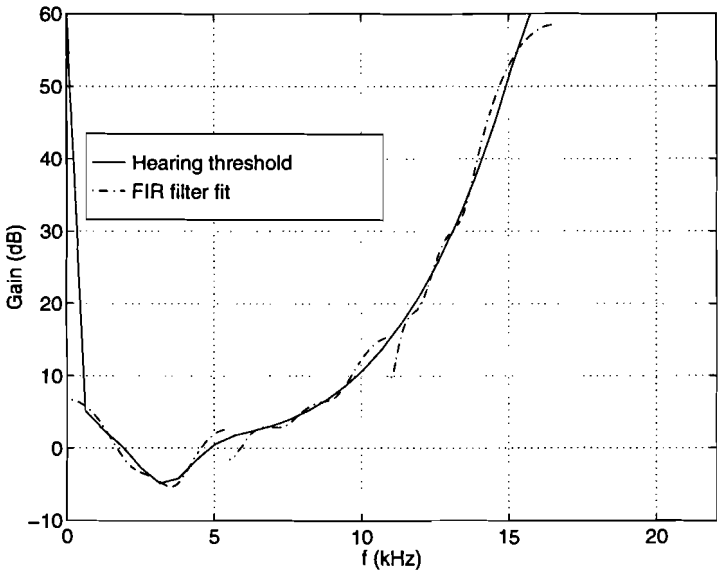


Figure 6.3: Hearing curve in quiet

Chapter 7

Transform Coding

The dithering technique can be applied to Transform Coding (TC), including the widely used Discrete Cosine Transform (DCT). In general the coding scheme is referred to as subband coding if the ratio $\frac{L}{M}$ is large, and it is referred to as transform coding if $\frac{L}{M}$ is small. The subband coding scheme from Chapter 6 reduces to a non-overlapping block transform if decimation factor M equals the FIR-filter length, L, of $\{H_0 \dots H_{M-1}\}$. An implementation of the dithered Transform Coding scheme is depicted in Figure 7.1. It can be proven that the transform matrices P and Q can be chosen such that this implementation is equivalent to the Subband Coding scheme of Figure 7.1.

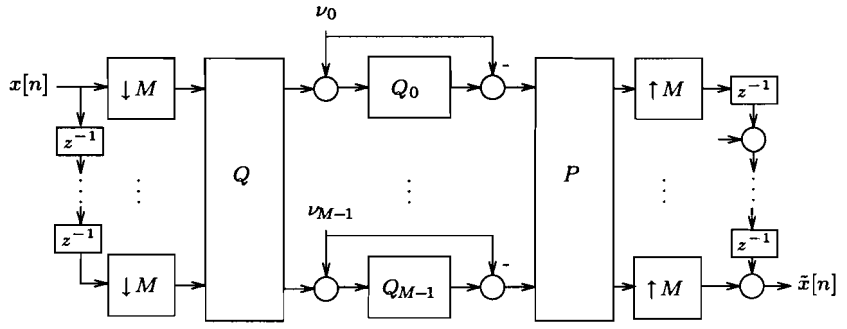


Figure 7.1: Implementation dithered Transform Coding scheme

In the following, the DCT-transform will be considered, combined with the usage of dither. This transform is very popular in the field of image processing. The forward 2-D DCT of an $n \times m$ block of the pixels $x(j,k)$ is defined as

$$X(u,v) = \frac{2C(u)C(v)}{\sqrt{n}\sqrt{m}} \sum_{j=0}^{n-1} \sum_{k=0}^{m-1} x(j,k) \cos\left[\frac{(2j+1)u\pi}{2n}\right] \cos\left[\frac{(2k+1)v\pi}{2m}\right], \qquad (7.1)$$

and the inverse 2-D DCT is defined as

$$x(j,k) = \frac{2}{\sqrt{n}\sqrt{m}} \sum_{j=0}^{n-1} \sum_{k=0}^{m-1} C(u)C(v)X(u,v) \cos\left[\frac{(2j+1)u\pi}{2n}\right] \cos\left[\frac{(2k+1)v\pi}{2m}\right], \qquad (7.2)$$

where

$$C(w) = \begin{cases} \frac{1}{\sqrt{2}}, & w = 0 \\ 1, & \text{otherwise} \end{cases} . \tag{7.3}$$

The transforms given by (7.1) and (7.2) are orthonormal. As a result, the error variance of the reconstructed image is equal to the sum of the error variances of the DCT coefficients. In most video applications, the block size is chosen to be 8×8 pixels. The image is first split into 8×8 non-overlapping blocks. Each block is then transformed independently. Perceptual weighting can be achieved by using a normalisation array that determines the relative quantisation stepsizes of all dct-coefficients. A typical normalisation array that has been used by JPEG in their studies is

$$\begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix} . \tag{7.4}$$

The high-frequency components have small energy. As a result, these coefficients are usually quantised to zero. The 8×8 blocks of quantised coefficients are ordered using a zigzag scan before encoding as is depicted in Figure 7.2.

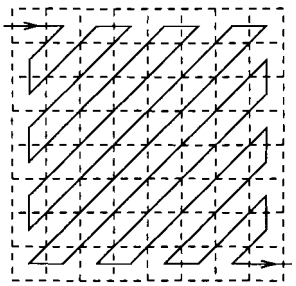


Figure 7.2: Zigzag scan of 8×8 DCT coefficients

In this way an 1-D sequence is obtained, in which the DCT coefficients are approximately ordered from high to low energy. The first coefficient in this sequence, the DC-coefficient is differentially coded, i.e. all DC coefficients are gathered, filtered with $\{1 - z^{-1}\}$ and entropy coded using a separate Huffman table. The reason for this is that the DC-coefficients are strongly correlated with each other. Intermediated zero runs in the remaining AC sequences are replaced by symbols indicating the number of zeroes (run-length coding). The last non-zero coefficients in the AC sequences are followed by an End-Of-Block (EOB) code. This EOB indicates that the rest of the 8×8 block merely contains zeroes.

The blocking artifacts introduced by the quantisation of the DCT-coefficients at low bit rates can be reduced by low pass filtering the edges of the blocks in the reconstructed image. The

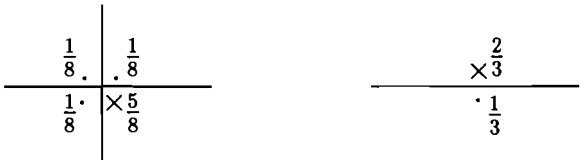


Figure 7.3: Low pass edge filtering

filtering is depicted in Figure 7.3. The pixels at the edges of the 8×8 blocks are processed as follows. If the pixel is at the corner of adjacent blocks, this pixel is multiplied by $\frac{5}{8}$ and added to the adjacent block pixels which are first multiplied by $\frac{1}{8}$. If the pixel to be processed is at the edge between two blocks only, it is multiplied by $\frac{2}{3}$ and added to the pixel at the opposite side of the edge which is first multiplied by $\frac{1}{3}$. In this way all pixels at the edges are processed. The remaining pixels are unaffected. In this way the intensity of the image is linearly interpolated at the edges so that the blocking decreases at the cost of a slight blur.

Chapter 8

Experimental audio results

In this chapter, experimental results for audio are presented for both wideband PCM and for subband coding.

8.1 Wideband PCM

Several dithering techniques were verified by means of audio experiments. In the experiments concerning wideband PCM, a system without any encoding is considered. In this way, experiments are performed fast and easily. 5 bits per sample were used to represent the quantised signal. The quantiser stepsize is thus determined by

$$\Delta = \frac{2A}{2^b - 1}, \tag{8.1}$$

where A is the maximum absolute amplitude of the audio sequence $x[k]$, i.e. $\max|x[k]|$, and $b = 5$ bit. Note that the entropy of the quantised signal will be lower than, or equal to 5 bit per sample, since the fixed length coding used does not exploit the signals probability distribution. Informal listening test where performed by the author only, and the subjective results are collected in Table 8.1 for the vocal excerpt 'witney'. The words 'subtractive' and 'non-subtractive' are abbreviated as 'S' and 'NS'. The results in the table are numbered and discussed in the following.

1. In the undithered case, the three different κ's yield a different noise colouring for input distributions of amplitudes comparable with the quantiser stepsize. κ represents the offset of the quantiser decision and representation levels and is defined by 1.4. The mid-tread quantiser ($\kappa = 0$) quantises to zero for input distributions below $\frac{1}{2}$LSB, while the mid-riser quantiser ($\kappa = \frac{\Delta}{2}$) maintains its noisy nature. The artifacts are most audible for a mid-tread quantiser, and can be characterised as gross signal dependent distortion.

2. If nonsubtractive RPDF dither is applied, the distortion disappears, but a clear signal dependent noise modulation remains. This noise modulation is observed as noise pumping, which is annoying.

Table 8.1: Listening test results

No.	Type of Dither	Subjective quality
1.	nondithered, $\kappa = 0, \frac{\Delta}{4}, \frac{\Delta}{2}$	clearly audible artifacts
2.	NS RPDF	signal dep. noise modulation
3.	NS TPDF	increase of noise floor, white error
4.	NS high pass	less perceptual noise power than 3.
5.	NS notch filtered	as 4. but less 'sharp' noise
6.	S RPDF	perfect, better than NS high pass
7.	S RPDF, 0.9LSB amplitude	terrible noise intermodulation
8.	S RPDF, 50% of dither randomly set to 0	reasonable
9.	S RPDF, even dither samples set to 0	artifacts
10.	S high pass	identical to S RPDF

3. This noise modulation is eliminated when TPDF dither is applied, at the cost of an increase of the noise floor. High pass dither results in less audible noise than nonsubtractive RPDF dither, although the total noise power is 50% higher.

4. The use of notch filtered dither, with coefficients {1 -1 1 -1 1} results in a less 'sharp' noise with perceptually equal noise power as for high pass dither in the case of coarse quantisation, although the noise power is twice as high than for nonsubtractive high-pass dither.

5. For 'fine' quantisation, say 7 bit per sample or more, high-pass dither is less audible than notch dither. This is a result of the loudness dependency of the psychoacoustical hearing curve.

6. RPDF subtractive dither of 1 LSB amplitude is undoubtly superior.

7. In an attempt to reduce the entropy, experiments have been done with subtractive RPDF dither of 0.9 LSB amplitude, as discussed in Section 3.1. The resulting signal still has all its undithered artifacts plus a noise compound due to the dither. Thus, the dithers amplitude must be set accurately to a multiple of the quantiser stepsize.

8. Setting the even dither samples to zero leads to a reduction of the artifacts discussed in item 1. The regular pattern with which the dither is set to zero is however annoying.

9. Setting a fraction of the dither samples to zero is a reasonable solution in between dithering and non-dithering, if it is done randomly.

10. The usage of subtractive dither always results in a noise power of $\frac{\Delta^2}{12}$, also if the dither was shaped with a FIR filter with integer coefficients. The summation of independent dither signals is also allowed. This is only important with respect to compatibility issues, i.e. several users use the coded data, but their are not all capable of subtracting the dither.

The best results are thus obtained by using RPDF subtractive dither. Highpass nonsubtractive dither can best be used if it is not possible to subtract the dither and 7 bit per sample or more are used to represent the audio sequence. Notch dither gives the best results for coarse quantisation (less than 7 bit per sample).

8.2 Subband Coding

In the following experiments, the vocal 'baby' fragment was used. Experiments where done with a four band subband system [12]. The quantiser stepsize was chosen to be equal in all subbands. These stepsizes can also be chosen in a perceptually more sophisticated way, as will be discussed in the next section. To give an indication of perceived quality, subjective remarks are included in Table 8.2. From the tests it follows that dithering band 0 and band 1

Table 8.2: SBC Listening test results

No.	Dithered Bands	H_0(bps)	Subjective quality
1.	0	0.5	artifacts
2.	1	0.5	modest artifacts
3.	2	0.5	no artifacts, much more noise than 2.
4.	0	1.0	artifacts
5.	1	1.0	modest artifacts
6.	2	1.0	no artifacts, more noise than 5.
7.	0	1.5	artifacts
8.	1	1.5	modest artifacts, BEST
9.	2	1.5	no artifacts, more noise than 8.

makes the error perceptually white, due to the low energy components in the higher subbands, at the cost of a significant increase in the noise floor in the second subband. Dithering band 0 only yields the best results for coarse quantisation. If higher bands that contain low energy are also dithered, the distortion increases as well as the entropy. The quantiser stepsize is then enlarged to obtain the desired bit rate, so that the distortion increases even further. Modest artifacts are still present however when dithering band 0 only. The artifacts mentioned in the table have the same nature as the artifacts in the undithered PCM signals and are very annoying.

8.3 Subband Coding and Noise Shaping

In this section, the quantiser stepsizes are chosen with the ratio's $\Delta_1 : \Delta_2 : \Delta_3 : \Delta_4 = 1 : 2.1 : 79.4 : 1000$, which correspond with the gains of 0dB, 6.4dB, 38dB and 60dB discussed in Section 6.2. Band 0 and band 1 of the audio fragment 'baby' where dithered and noise shaped. The resulting noise power is depicted in Figure 8.1 for the case that no noise shaping was used (a), and for the case that the noise shaping filters of Section 6.2 where used in band 0 and band 1 (b). From Figure 8.1 (a) follows that the noise floors are flat and 6.4 dB apart in band 0 and band 1. The noise power is much smaller in the third and fourth band, since

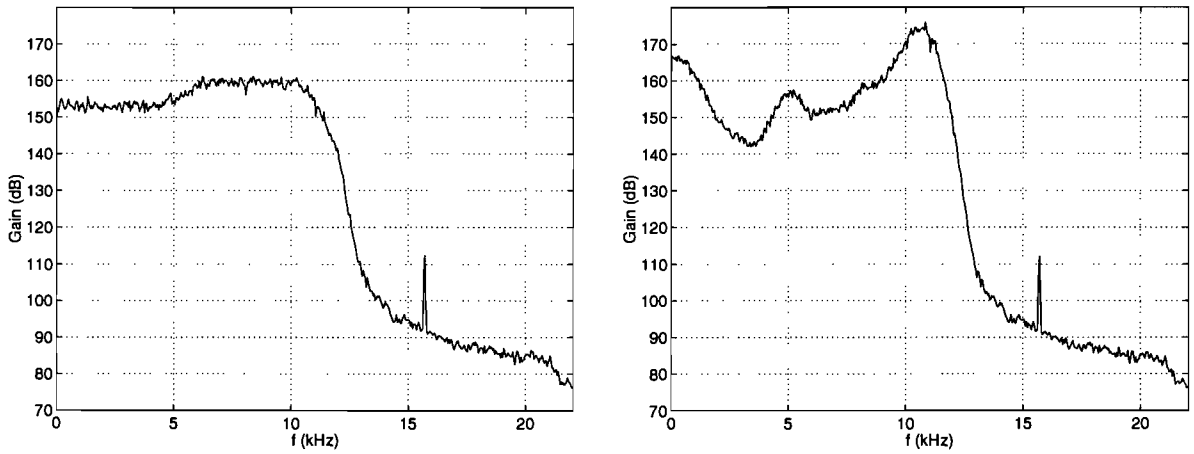


Figure 8.1: Quantisation without noise shaping (a), and noise shaped quantisation (b).

the signal power is smaller than $\frac{\Delta^2}{12}$ in these bands. In Figure 8.1 (b) the noise in band 0 & 1 is shaped according to Figure 6.3. The peak in the spectrum at 15.6kHz is due to television and video interference[1]. Some results of the listening tests involving dither and noise shaping are gathered in Table 8.3. Dithering and noise shaping band 0 introduces much noise at a

Table 8.3: Listening test results for SBC with Noise Shaping

Dithered bands	Shaped bands	H_0(bps)	Comments
1	1	0.5	much noise, artifacts decrease
2	2	0.5	0.5 bps cannot be achieved
1	1	1.0	audible noise, better than unshaped
2	2	1.0	much & sharp noise
0	1	1.5	Somewhat less artifacts due to noise shaping
0	2	1.5	same as shaping band 0 only
1	1	1.5	noise near threshold
2	2	1.5	audible noise

bit rate of 0.5 bit per sample (bps), artifacts decrease however. This experiment also showed that the entropy of a dithered and noise shaped subband approximately equals 1 bps for large quantiser stepsizes. For this reason, dithering and noise shaping band 0 & 1 is not possible at 0.5 bps without a phenomenal increase in distortion. At 1 bps, dithering and noise shaping band 0 only gives the best results. The same holds for 1.5 bps. At this bit rate the noise approaches the hearing threshold for the excerpt used. Listening test where also done at bit rates higher than 2.0 bps. These tests pointed out that the use of dither gives improvement for bit rates of 2 bps or lower. This is because quantisation error become less correlated at higher bit rates, so that artifacts disappear. Only the subbands that contain an amount of

[1]Video equipment uses 25 pictures per second of 625 lines (585 are visible) conform the CCIR norm. The generated frequency component is present at many compact disc and equals 625 × 25 Hz = 15.625kHz.

energy of $\frac{\Delta^2}{12}$ or more should be dithered and noise shaped. This is because dithering and noise shaping the low energy bands results in an excessive increase in noise power and bit rate. Dithering band 0 only, gives better results than dithering band 0 and band 1. A good quality can be achieved at 1.5 bit for the excerpt used. The listening tests also showed that noise shaping itself reduces the signal artifacts, even when no dither is applied. This is because the highly correlated quantisation noise in an undithered system is diffused by the noise shaping. Signal artifacts due to the undithered quantisation of subband 2 where present at some bit rates in the form of impulses. Choosing the quantiser stepsizes a factor of two smaller in the undithered bands with respect to the ratios used, can be a solution to this problem.

Chapter 9

Experimental video results

In this chapter, simulation results for video are presented for both plain quantisation and quantisation of DCT coefficients in a transform coding scheme. Plain quantisation is performed on the 'Teeny' image, while transform coding is applied to the 'Lenna' image. Reconstructed images are included, since a lower mean square error does not necessarily imply a perceptually better image quality. Furthermore, the perceptive quality depends highly on the image resolution and the viewing conditions like for example the viewing distance.

9.1 Image quality

The image quality is measured using the Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR). Denoting the original $n \times m$ image by x and the reconstructed image by $\tilde{x}$, the MSE is given by

$$\mathrm{MSE} = \frac{1}{nm} \sum_{j=0}^{n-1} \sum_{k=0}^{m-1} [x(j,k) - \tilde{x}(j,k)]^2 \,. \tag{9.1}$$

The PSNR is calculated from the MSE using

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{255^2}{\mathrm{MSE}} \right) \,(\mathrm{dB})\,, \tag{9.2}$$

for an 8-bit image. Error images, representing the difference between the original and reconstructed images, are also shown. The error image, e, is generated using

$$e(i,j) = 2\,[x(j,k) - \tilde{x}(j,k)] + 128\,. \tag{9.3}$$

The error image has been magnified by a factor 2, so that possible artifacts become clearly visible. Furthermore it has been biased by 128 so that the error can be displayed on an output device. The error image of a perfectly reconstructed image is a gray field of value 128, while the error image of a perfectly dithered image consists of a white noise component around a mean of 128.

9.2 Plain quantisation

In this section, results are presented for simple PCM systems which output is entropy coded and in which the dithering technique is used. The test image 'Teeny' is quantised using no dither, subtractive and non-subtractive RPDF dither, and non-subtractive TPDF dither at a rate of 2.0 bit per pixel (bpp). The resulting images and the residual images are included in Appendix B. The corresponding PSNR values are collected in Table 9.1. Subtractive and non-subtractive dither are indicated with 'SD' and 'NSD' respectively. In the table, subtractive TPDF dither is also considered. This type of dither can be useful if there are multiple users of the encoded images, of which only some are able to subtract the TPDF dither.

Table 9.1: Results for Teeny, H_0=2.0 bbp

Dither	Δ	PSNR(dB)	Wiener PSNR(dB)	H_1(bpp)
No Dither	58	23.3	24.0	0.42
RPDF NSD	61	20.0	27.7	1.20
RDPF SD	"	23.2	31.0	"
TPDF NSD	66	17.8	25.3	1.78
TPDF SD	"	22.5	30.4	"

The 8-bit original image is quantised in approximately 4 output levels in the undithered case. Since the distribution of 'Teeny' is approximately uniform, the bit rate of 2.0 bbp is in agreement with Figure 2.3 on page 13 which indicates the bit rate of a uniformly distributed quantised signal. Furthermore, experiments indicated that the entropy of the undithered quantised image varies with tenths of bits as κ is varied, which is also in agreement with the figure mentioned. This κ-dependency is only important for undithered coarse quantisation. Note that the resulting PSNR of the RPDF subtractively dithered image approximately equals the PSNR of the undithered quantised image, although the subjective image quality has improved using the dither. The noise introduced by the dithered quantisation process is easier to remove using noise-reduction techniques like Wiener filtering than the highly correlated quantisation noise generated by an undithered quantiser. The Wiener filtered subtractively RPDF dithered image is also included in Appendix B. The PSNR is improved by more than 7 dB. The PSNR of the Wiener filtered images are also included in Table 9.1.

The entropy discussed so far, is generally referred to as the zero-order entropy. Higher order entropies are based on conditional probabilities rather then on zero order statistics. When the first order entropy is calculated, the statistics of the pixels are considered given the neighbouring pixel values. The dithering technique decorrelates the adjacent pixels, so that higher order entropies are much higher in the case that the entropy coder has no knowledge of the dither. This is illustrated by the images first order entropy, H_1, which is also included in Table 9.1. The complexity of the encoder and decoder increase exponentially with the entropy order. As a result, a zero-order entropy coding and run-length coding are most commonly used.

9.3 Transform coding

The DCT-transform will be used in the experiments discussed in this section. The resulting images are included in Appendix C. RPDF dither was used in all experiments in combination with a deadzone quantiser. In the results presented, the optimal deadzone was used with respect to the perceptual quality. The subjective results are summed in Table 9.2. All images presented except for the original image are quantised at 0.5 bpp. The entries in the table are

Table 9.2: Results for Lenna, H_0=0.5 bbp

	Dither	zone (Δ)	weighting	PSNR(dB)	Remarks
1.	-	-	-	-	Original
2.	N	1.5	N	30.49	-
3.	N	1.5	N	30.44	2, low pass edge-filtered
4.	N	1.5	Y	29.88	
5.	N	1.5	Y	29.57	4, Wiener filtered
6.	Y	1.0	N	15.33	-
7.	Y	3.5	N	28.57	-
8.	Y	3.5	N	30.00	7, Wiener filtered
9.	Y	3.5	N	30.18	7, no zero subtraction
10.	Y	3.0	Y	19.80	-
11.	Y	3.0	Y	26.89	10, Wiener filtered
12.	Y	3.0	Y	29.61	10, no zero subtraction
13.	N	1.5	N	-	2, residual
14.	Y	3.5	N	-	9, residual
15.	N	1.5	Y	-	4, magnified
16.	Y	3.0	Y	-	12, magnified

numbered and discussed in the sequel.

1. Original 8-bit Lenna image.

2. Uniform quantisation of the DCT-coefficients without the use of dither or frequency weighting. This image shows clear blocking artifacts, especially around the shoulder, and ringing artifacts, especially round the edges of the hat. These artifacts are clearly visible in the residual image which is included in image 13.

3. The blocking is partially eliminated by the low pass edge filtering of image 2, discussed in Chapter 7, at the cost of a slight blur.

4. Also no dither was used, but the weighting array (7.4) on page 40 was used. The perceptual quality of this image is much better than that of image 2, despite the decrease in PSNR of 0.92dB. Blocking and ringing are still present however.

5. Another way of reducing the signal artifacts is by using a Wiener filter. This image is the Wiener filtered version of image 4. Clearly, Wiener filtering also introduces a

significant blur. Image 15 is a magnification of this image, and clearly displays that the Wiener filter blurs the edges of the image.

6. All DCT-coefficients are dithered. The higher frequency components in this image are no longer quantised to zero. As a result, the quantiser stepsize was chosen very large in order to reduce the entropy to 0.5 bpp. In addition, the higher frequency components usually have less energy than $\frac{\Delta^2}{12}$, so that the noise power increases due to the use of dither.

7. In order to overcome this problem, a deadzone quantiser is used in the following images. The best deadzone is 3.5Δ if no weighting is used. This is in agreement with Figure 4.7 on page 28. The image still shows blocking and ringing.

8. This is image 7, Wiener filtered. This image does not look better than the undithered Wiener filtered image 5. The dithered version could possibly be improved a little more if frequency weighting was used.

9. This image was processed like image 7, except that the dither was not subtracted if the quantiser output equals zero. The image looks less noisy than image 7, but has comparable artifacts.

10. This image was processed using frequency weighting and dither. The optimal deadzone is 3Δ if weighting is used. The noise in this image has a larger variance than in the unweighted image 7.

11. Wiener filtering image 10 results in 'spots' randomly distributed over the image. This is a result of the large noise variance introduced by the dither.

12. Weighting, dithering and not subtracting the dither if the quantiser output is zero, yields image 12. This image also shows no significant improvement caused by the use of dither. The artifacts are comparable to those of the undithered image 4.

13. Residual of image 2.

14. Residual of image 9.

15. Image 4, magnified.

16. Image 12, magnified.

The images included show that using dither and deadzone quantisation in a DCT-system does not improved the PSNR or the perceptual image quality. Furthermore, experiments where done with uniform quantisers, where only the 6 lowest frequency coefficients out of 64 were dithered. This also showed no improvement. Our results show that dither only results in perceptual improvement in systems where highly correlated signals are quantised such as in PCM systems.

Chapter 10

Summary

In a lossy data compression system, the resulting error signal can be highly correlated with the input. This noise can be perceptually annoying. It can be made white and independent of the input signal by subtractively dithering the quantisers incorporated in the data compression system. The noise power can increase excessively due to the dither when using the same bit rate. It is of interest to know if the usage of dither can improve the resulting perceptual quality. Also the possibility of using dither in existing systems is important.

It is shown that dither can be used within an existing implementation of a noise shaped quantiser. This can be important when an existing noise shaped quantiser is used at low bit rates and produces signal artifacts and limit cycles.

A theoretical rate-distortion analysis was performed on uniform and generalised Gaussian sources. Also the usage of deadzone quantisers was considered. The optimal deadzone width was calculated from the rate-distortion curves. Experiments where performed with dither in simple PCM systems for both audio and video. Improvements are reported for low bit rate applications.

Dither was also incorporated in a four-band audio subband coding system. Artifacts can be reduced when dithering the lowest band only. The white noise introduced by the dithered quantiser can successfully be noise shaped. At a bit rate of 2 bit per sample or more, the error sequence already has a random nature without the usage of dither. At bit rates lower than 2 bpp, perceptual improvements are reported.

The use of dither was also investigated for a video DCT system. Deadzone quantisers where used to reduce the entropy. Experimental results are included. No improvement was obtained using the dither.

Bibliography

[1] N.S. Jayant and P. Noll,
Digital Coding of Waveforms
Englewood Cliffs, New Jersey: Prentice-Hall, 1984

[2] Lipshitz S.P. et al,
Quantization and Dither: A Theoretical Survey
J. Audio Eng. Soc., vol. 40, pp. 355-375, May 1992

[3] Widrow, B.,
A Study of Rough Amplitude Quantisation by Means of Nyquist Sampling Theory
IRE Trans. Circuit Theory, vol. 3, pp. 266-276, Dec. 1956

[4] Roberts, L.G.,
Picture Coding Using Pseudo-Random Noise
IRE Trans. Information Theory, vol. 8, pp. 145-154, Feb. 1962

[5] Schuchman, L.,
Dither Signals and Their Effect on Quantization Noise
IEEE Trans. Commun. Tech., vol. 12, pp. 162-165, Dec. 1964

[6] Sherwood D.T.,
Some Theorems on Quantization and An Example Using Dither
Conference Record, 19th Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, Nov. 1985.

[7] Gerzon A.M. et al,
Psychoacoustic Noise Shaped Improvements in CD and Other Linear Digital Media
Proceedings of the 94th Convention of the Audio Eng. Soc., Berlin, Germany, March 1993

[8] Lipshitz S.P. et al,
Dithered Noise Shapers and Recursive Digital Filters

[9] R.G. Gallager,
Information Theory and Reliable Communication
Wiley, New York, 1964

[10] T. Chen,
Elimination of Subband-Coding Artifacts using the Dithering Technique
IEEE Proc. ICIP vol. II, pp. 874-877, 1994

[11] K. Sharifi and A. Leon-Garcia,
Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video
IEEE Trans. Circuits Syst. Video Techn., vol. 5, pp52-56, Feb. 1995

[12] M.E. Groenewegen,
A low complexity Hi-Fi audio codec
Nat. Lab. Technical Note Nr. 327/94, Okt. 1994

[13] M. Rabbani and P.W. Jones,
Digital Image Compression Techniques
SPIE Optical Engineering Press, Washington, January 1991

List of Symbols and Abbreviations

$\forall$	for all
$\lfloor x \rfloor$	Largerst integer less than or equal to x
$\lceil x \rceil$	Smallest integer greater than or equal to x
$\triangleq$	Equality by defenition
$\mathcal{B}[\Gamma]$	$\mathcal{B}$ operator, returns $-\Gamma \log_2(\Gamma)$
Δ	Quantiser stepsize
ϵ	Quantising error
κ	Offset quantising decision and representation levels
ν	The dither signal
$\Pi_\Gamma(x)$	rectangular window function, returns $\frac{1}{\Gamma}$ if $-\frac{\Gamma}{2} < x \leq \frac{\Gamma}{2}$, and 0 otherwise
$\mathbb{Z}^N$	The N-dimensional set of integers
$E[f]$	Expection, returns $\int_{-\infty}^{\infty} f(\epsilon) p_\epsilon(\epsilon) d\epsilon$
$H(.)$	Entropy
$I(.)$	Information
LSB	Corresponds with Δ
$p_x(x)$	Probability density of x
$p_{x_1,x_2}(x_1, x_2)$	Joint pdf of x_1 and x_2
$P_x(x)$	Characteristic function of x
$P_{x_1,x_2}(x_1, x_2)$	Joint cf of x_1 and x_2
u	The fourier variable corresponding with x
x	The system input signal
$\text{sinc}(x)$	$\frac{\sin(x)}{x}$
$\tilde{x}$	The system output signal
bpp	Bit per Pixel
cf	Characteristic Funtion, i.e. fourier transform of a pdf
pdf	Probability Density Function
PSD	Power Spectral Density
RPDF	Rectangular Probability Density Function
TPDF	Triangular Probability Density Function
mRPDF	Pdf formed by m-fold convolution of RPDF

Appendix A

The theory of dithering

This appendix provides a survey of fundamental theorems concerning subtractively and non-subtractively dithered systems as well as non-dithered systems. A derivation of these theorems can be found in [2]. Practical implications that follow from these theorems are included as remarks.

Theorem 1 *The total error induced by an undithered quantising system is uniformly distributed if and only if the cf of the system input, P_x, satisfies the condition that*

$$P_x(u)\Big|_{u=\frac{k}{\Delta}} = 0 \qquad \text{for } k \in \mathbb{Z} \setminus \{0\}\,. \tag{A.1}$$

Remarks: The uniform distribution of the error means that the error signal is an RPDF process. It does not imply however that the error signal is white and independent of the input signal; in an undithered quantising system, the quantisation error is deterministically related to the systems input.

Theorem 2 *In an undithered quantising system, the joint pdf of total error values ϵ_1 and ϵ_2, separated in time by $\tau \neq 0$, is given by*

$$p_{\epsilon_1,\epsilon_2}(\epsilon_1,\epsilon_2) = \Pi_\Delta(\epsilon_1)\Pi_\Delta(\epsilon_2)\,, \tag{A.2}$$

if and only if the joint cf, P_{x_1,x_2}, of the corresponding system inputs, x_1 and x_2, obeys the condition that

$$P_{x_1,x_2}(u_1,u_2)\Big|_{(u_1,u_2)=(\frac{k_1}{\Delta},\frac{k_2}{\Delta})} = 0 \qquad \text{for } k_1,k_2 \in \mathbb{Z}^2 \setminus \{0,0\}\,. \tag{A.3}$$

Remarks: The joint pdf is of interest, since it determines the power spectral density of the error signal. If and only if (A.3) is satisfied, the joint pdf is uniformly distributed conform (A.2). Again, this does not imply that the error signal is independent of the quantising systems input. This error sequence itself is white however.

Theorem 3 (Widrow's Quantising Theorem) *The pdf of the input to an undithered infinite linear quantising system is recoverable from the pdf of its output if the cf of the input, P_x, is band limited such that $P_x(u) = 0$ for $|u| \geq \frac{1}{2\Delta}$.*

Remarks: Widrow's quantising theorem is analogous to Nyquist's well known theorem for time sampling. It means that the pdf of the input signal can faithfully be reconstructed from the quantised signal if the characteristic function of the input signal is properly band limited.

Theorem 4 (Schuchman's Condition) *In a subtractively dithered quantising system, the pdf of the total error will be uniformly distributed and statistically independent of the input for arbitrary input distributions if and only if the cf of the dither, P_ν, satisfies the condition that*

$$P_\nu(u)\Big|_{u=\frac{k}{\Delta}} = 0 \qquad \text{for } k \in \mathbb{Z} \setminus \{0\}\,. \tag{A.4}$$

Remarks: Any dither with a uniform distribution of 1 LSB top-to-top amplitude satisfies this theorem. (e.g. a saw-tooth dither of 1 LSB top-to-top amplitude). The theorem does not state anything about the power spectral density of the total error. The total error itself is statistically independent of the system input, but is not necessarily white due to possible correlation in the dither.

Theorem 5 *In a subtractively dithered quantising system, the joint cf, $P_{\epsilon_1,\epsilon_2}$, of two total error values, ϵ_1 and ϵ_2, (separated in time by $\tau \neq 0$) is independent of the joint cf, P_{x_1,x_2}, of the corresponding input values, x_1 and x_2, and is given by*

$$P_{\epsilon_1,\epsilon_2}(u_1, u_2) = sinc(\pi\Delta u_1)sinc(\pi\Delta u_2)\,, \tag{A.5}$$

for arbitrary input distributions if and only if

$$P_{\nu_1,\nu_2}(u_1, u_2)\Big|_{(u_1,u_2)=(\frac{k_1}{\Delta},\frac{k_2}{\Delta})} = 0 \qquad \text{for } k_1, k_2 \in \mathbb{Z}^2 \setminus \{0, 0\}\,. \tag{A.6}$$

Remarks: If the conditions of this theorem are satisfied, then ϵ_1 and ϵ_2 are both uniformly distributed and statistically independent of one another. Note that if ν_1 and ν_2 are statistically independent of one another, and the cf of each satisfies (A.4), then (A.6) will be satisfied. Clearly, the simplest dither obeying (A.4) is a white RPDF signal of 1 LSB top-to-top amplitude. But also a summation of independent RPDF signals with integer multiples of 1 LSB amplitude obeys (A.4).

Theorem 6 *In a non-subtractively dithered quantising system, $E[\epsilon^m|x]$ is functionally independent of x if and only if*

$$\frac{d^m G_\nu}{du_x^m}(u)\Big|_{u=\frac{k}{\Delta}} = 0, \qquad \text{for } k \in \mathbb{Z} \setminus \{0\}\,, \tag{A.7}$$

where

$$G_\nu(u) \triangleq sinc(\pi\Delta u)P_\nu(u)\,. \tag{A.8}$$

Remarks; The m^{th} error moment is functionally independent of x when (A.7) is satisfied.

Theorem 7 *In a non-subtractively dithered quantising system, $E[\epsilon^m|x]$ is functionally independent of x for $m = 1, 2, ..., M$ if and only if*

$$\frac{d^i P_v(u_x)}{du_x^i}(u_x)\Big|_{u_x=\frac{k}{\Delta}} = 0, \qquad \text{for } k \in \mathbb{Z} \setminus \{0\} \text{ and } i = 0, 2, ..., M-1\,. \tag{A.9}$$

Remarks: All first M error moments are functionally independent of x when this theorem is satisfied.

Theorem 8 *A non-subtractive dither signal generated by the summation of n independent rectangular-pdf random processes, each of 1 LSB peak-to-peak amplitude, renders the first n moments of the total error independent of the system input, and results in a total error power of $\frac{(n+1)}{12}\Delta^2$.*

Remarks: This theorem provides practical dither signals that satisfy (A.9).

Appendix B

Plain quantised video results

Figure B.1: Original image (a), Quantised image at 2 bpp (b), RPDF non-subtractively dithered at 2 bpp (c), TPDF non-subtractively dithered at 2 bpp (d), RPDF subtractively dithered at 2 bpp (e) and RPDF subtractively dithered Wiener filtered at 2 bpp (f).

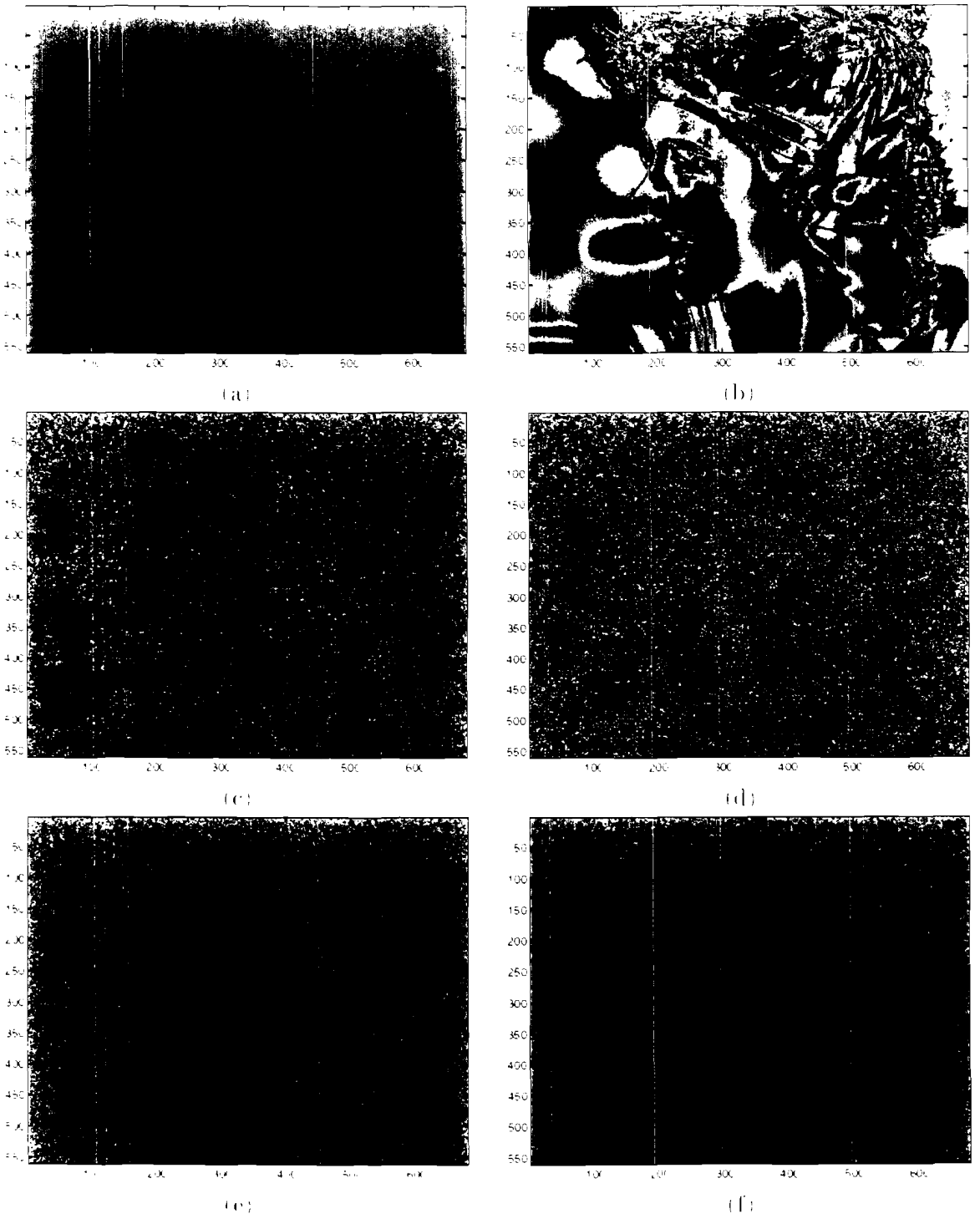
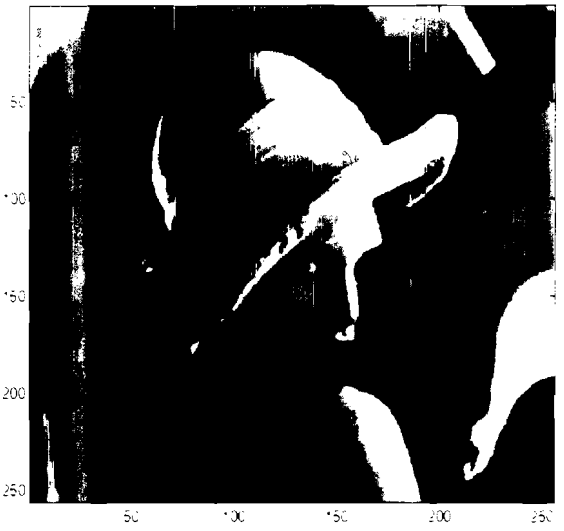


Figure B.2: Residuals of the images in Figure B.1.

Appendix C

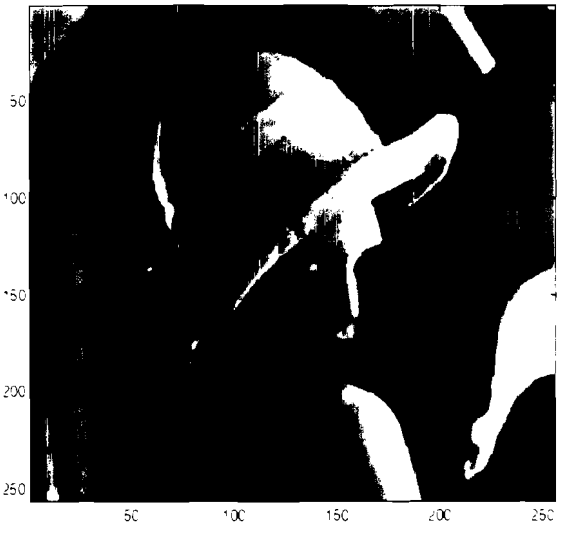
DCT video results

1. Original



2. Undithered

3. Undithered, low-pass edge filtered

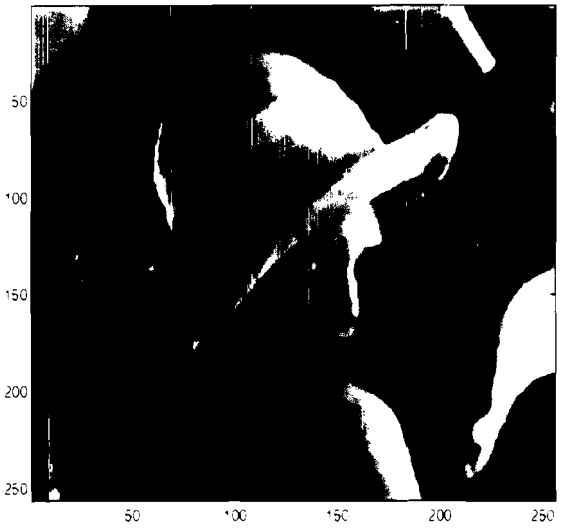


4. Undithered, weighted

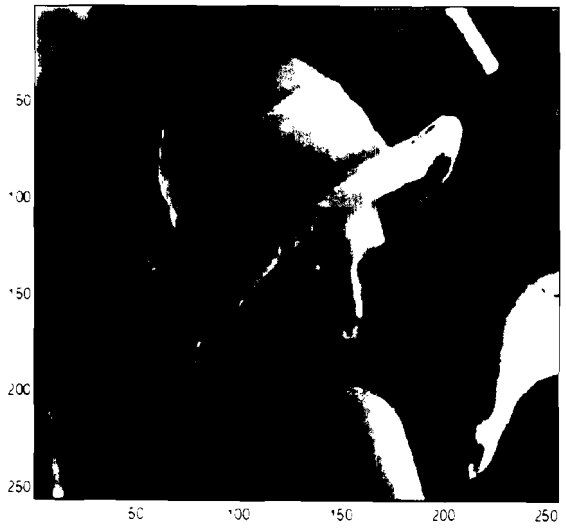
5. Undithered, weighted, Wiener filtered

6. Dithered, using a uniform quantiser

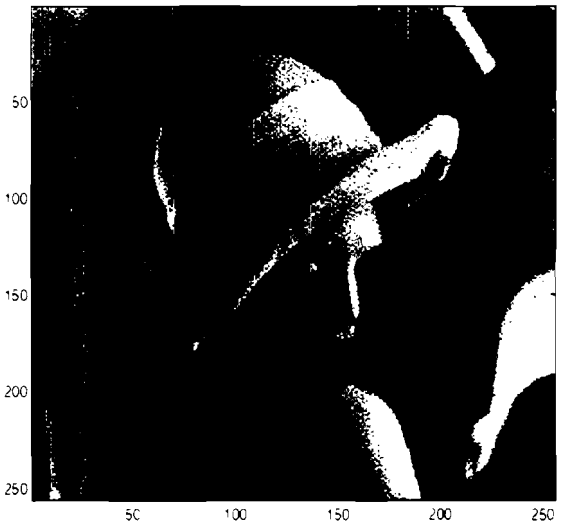
7. Dithered



8. Dithered, Wiener filtered

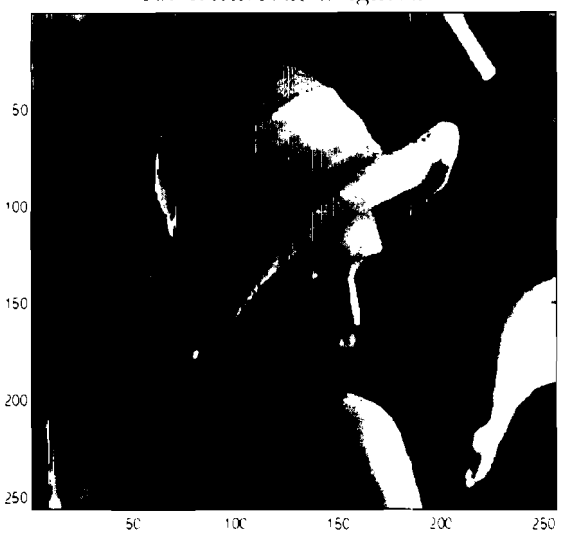


9. Dithered, no zero subtraction

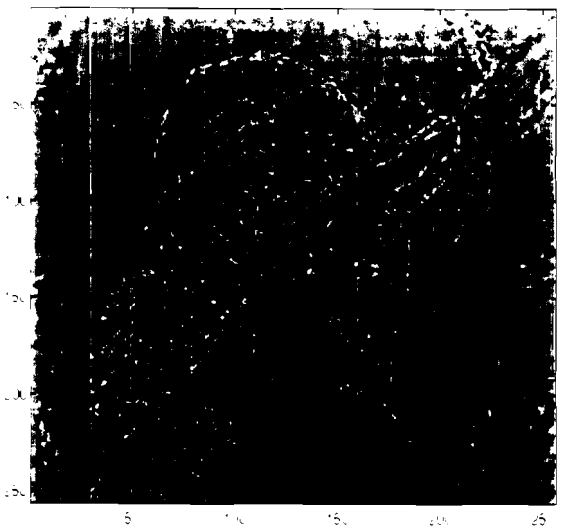


10. Dithered, weighted

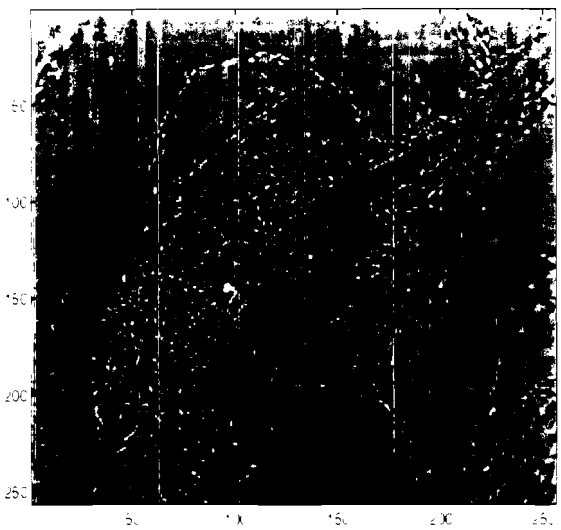
11. Dithered, weighted, Wiener filtered



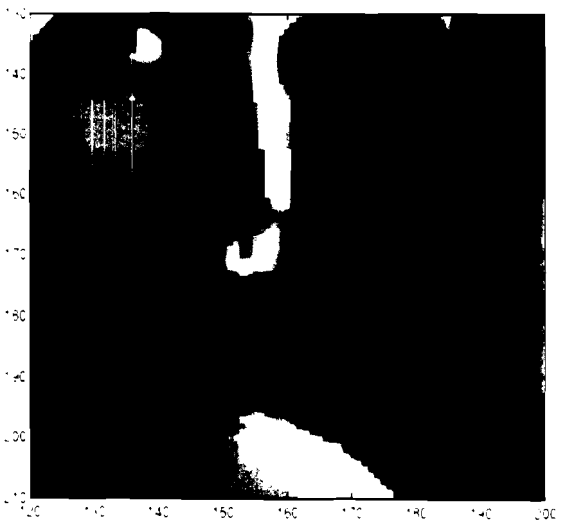
12. Dithered, weighted, no zero subtraction



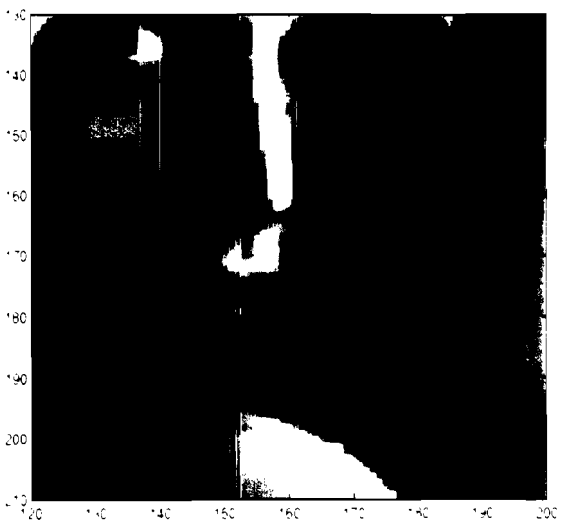
13. Residual of picture 2



14. Residual of picture 9



15. Picture 1, magnified



16. Picture 12, magnified