ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ:
«Ανάλυση θορύβου και ευστάθειας σε μετατροπείς υπερδειγματοληψίας»

ΦΟΙΤΗΤΗΣ : Δασκαλάκης Χαράλαμπος
ΕΠΙΒΛΕΠΩΝ : Ευθυβουλίδης Γεώργιος

Βόλος, Μάιος 08

Πρόλογος

Η παρούσα εργασία έχει στόχο να παρουσιάσει βασικές λειτουργικές πτυχές, μιας από τις αρχιτεκτονικές μετατροπέων αναλογικού σε ψηφιακό σήμα (και αντίστροφα), που βασίζονται στην υπερδειγματοληψία. Η προετοιμασία της ξεκίνησε τον Σεπτέμβριο 2007 και ολοκληρώθηκε το Μάιο 2008. Το έναυσμα δόθηκε από το προπτυχιακό μάθημα επιλογής: Σχεδίαση Αναλογικών Κυκλωμάτων VLSI, όπου διδάχτηκαν και οι βάσεις για το συγκεκριμένο θέμα.

Ως βάση της ανάλυσης χρησιμοποιήθηκε ένα από τα λίγα βιβλία [13] που έχουν γραφτεί για αυτό το θέμα, το οποίο συγκεντρώνει μεγάλο μέρος των δημοσιεύσεων που έχουν γίνει μέχρι την χρονολογία έκδοσης του (2004).

Παρότι σήμερα υπάρχουν πολλά προϊόντα σε αυτό το χώρο, η σχεδίαση τους δεν βασίζεται τόσο σε πλήρη θεωρητική τεκμηρίωση, όσο σε πειραματικές προσομοιώσεις. Αυτό συμβαίνει γιατί η μη γραμμική λειτουργία των βροχών ανάδρασης, δημιουργεί απρόβλεπτη σε ένα βαθμό συμπεριφορά, και καθιστά αδύνατη τη μοντελοποίησή της με τα υπάρχοντα μαθηματικά εργαλεία.

Αρχικά μελετήσαμε θεωρητικά και επιβεβαιώσαμε με προσομοιώσεις τις βασικές αρχές λειτουργίας των μετατροπέων υπερδειγματοληψίας, συμπεριλαμβανομένης και της ανάλυσης θορύβου καθώς και την επίδραση ορισμένων φίλτρων σε αυτόν. Έπειτα βασιζόμενοι σε ορισμένα πορίσματα της ανάλυσης ευστάθειας, μελετήσαμε τα όρια της ευστάθειας για διάφορα σήματα και περιπτώσεις. Τα αποτελέσματα είναι πειραματικής φύσεως, καθώς όπως είπαμε ο τρόπος με τον οποίο το σύστημα τελικά οδηγείται σε αστάθεια παραμένει ανεξήγητος.

Καθότι δεν υπάρχει βιβλιογραφία για το συγκεκριμένο αντικείμενο στην ελληνική γλώσσα, προς αποφυγή άστοχων μεταφράσεων της τεχνικής ορολογίας, χρησιμοποιήθηκε η αγγλική γλώσσα στο σύνολο της εργασίας, που εμπεριέχεται ως παράρτημα.

Μετατροπείς Υπερδειγματοληψίας

Για να μετατρέψουμε ένα σήμα από αναλογική σε ψηφιακή μορφή, συχνά επιδιώκουμε μεγάλη ακρίβεια. Αυτή η ακρίβεια μετριέται σε bits και κυμαίνεται στις μέρες μας από 4 έως 22 bits ανάλογα με τις απαιτήσεις της κάθε εφαρμογής. Οι μετατροπείς με ακρίβεια παραπάνω από 12 bits συχνά χρησιμοποιούν μεθόδους υπερδειγματοληψίας. Σύμφωνα με το θεώρημα του Nyquist απαιτούνται τουλάχιστο δύο δείγματα ανά περίοδο. Ο λόγος υπερδειγματοληψίας δείχνει πόσες φορές πάνω από αυτό το όριο δειγματοληπτούμε το σήμα. Όσο μεγαλύτερος είναι ο λόγος αυτός τόσο μεγαλύτερη ακρίβεια πετυχαίνουμε. Έτσι για παράδειγμα εάν δειγματοληπτήσουμε ένα ημιτονικό σήμα συχνότητας 1 KHz με λόγο υπερδειγματοληψίας 128, τότε θα πρέπει να έχουμε 256 δείγματα ανά περίοδο του σήματος. Έτσι μπορούμε με χρήση ενός απλού κβαντιστή του ενός bit συνδεδεμένο σε κατάλληλο βρόχο ανάδρασης να πετύχουμε με κατάλληλο λόγο υπερδειγματοληψίας ακρίβεια μετατροπής μέχρι και 10 bits.

Για να καταλάβουμε καλύτερα σε ποιες τιμές κυμαίνονται οι μετατροπείς αναλογικού σε ψηφιακό σήμα, θα πρέπει να δούμε χωριστά τις επιδόσεις της κάθε κατηγορίας. Οι μετατροπεις που δεν χρησιμοποιούν υπερδειγματοληψία, αλλα δειγματοληπτούν σε συχνότητα Nyquist, μπορούν να φτάσουν μέχρι και 20 Gs/s (20.000.000.000 δείγματα το δευτερόλεπτο) με μόλις 6.5 bits ακρίβεια [6], εις βάρος όμως της κατανάλωσης ισχύος που φτάνει τα 10 W. Άλλες σχεδιάσεις, με λιγότερο απο 1 W κατανάλωση, φτάνουν τα 1 Gs/s με 8.85 bits ακρίβεια [7], ή 0.8 Gs/s με 9 bits [5], ή 1.35 Gs/s με 7.7 bits [9] ή με βελτίωση του προηγουμενου σχεδίου μέχρι 1.8 Gs/s με 8.3 bits [15]. Άπο την άλλη μεριά, οι μετατροπείς υπερδειγματοληψίας μπορούν να φτάσουν μέχρι και 20 ή και 22 bits, με πολύ αργό ομως σήμα εισόδου: 15 ή 12.5 Hz αντίστοιχα [11], [10]. Ποιο ισορροπημένες σχεδίασεις που περιγράφονται στο [12], έχουν ακρίβεια γύρω στα 15 bits με συχνότητα σήματος μέχρι 1 MHz.

Στους μετατροπείς υπερδειγματοληψίας, το πλήθος των βρόχων ανάδρασης ορίζει το βαθμό του μετατροπέα. Έτσι για πρώτου βαθμού μετατροπέα συναντάμε ένα βρόχο ανάδρασης, για δεύτερου βαθμού δύο, και αντίστοιχα για μεγαλύτερο βαθμό. Βέβαια η διάταξη του κάθε βρόχου, καθώς και το κέρδος του καθορίζουν το τελικό αποτέλεσμα.

Είναι εύκολα κατανοητό ότι η υπερδειγματοληψία μπορεί να εφαρμοστεί σε σχετικά αργά σήματα, όπως τα ακουστικά, καθώς για παράδειγμα ένα σήμα συχνότητας 100 MHz με λόγο υπερδειγματοληψίας 128 θα απαιτούσε 12.8 GHz συχνότητα δειγματοληψίας που είναι πολύ δύσκολο να επιτευχθεί.

Ο κβαντιστής στην πιο απλή του μορφή μπορεί να είναι ένας συγκριτής με δύο στάθμες εξόδου. Μπορούμε εναλλακτικά να χρησιμοποιήσουμε ένα μετατροπέα με περισσότερες στάθμες εξόδου. Συχνά αυτός ο μετατροπέας είναι ένας Nyquist μετατροπέας λίγων bits που χρησιμοποιείται σε συνδεσμολογία υπερδειγματοληψίας προκειμένου να βελτιωθεί η ανάλυση του.

Έτσι για παράδειγμα εάν έχουμε ένα σήμα εισόδου $0.25\,\mathrm{V}$ και τροφοδοτούμε έναν μετατροπέα που χρησιμοποιεί έναν κβαντιστή με επίπεδα 0 και 1 σε μονό βρόχο ανάδρασης, τότε θα αναμένουμε στην έξοδο του κβαντιστή να έχουμε έξοδο της μορφής 1 0 0 0 1 0 0 0 1 0 0 0 κοκ. Η μέση τιμή της εξόδου παρατηρούμε ότι μετά από ικανό πλήθος δειγμάτων προσεγγίζει την τιμή της εισόδου.

Ιδιαίτερη σημασία θα πρέπει να δοθεί στον θόρυβο κβαντοποίησης που παραμένει κοντά στη συχνότητα του σήματος. Η μοντελοποίηση του ως τυχαίο σήμα, μας δίνει κάποιες προσεγγιστικές εκτιμήσεις για το μέγεθος του, καθώς και τη συσχέτιση του με το λόγω υπερδειγματοληψίας και την αρχιτεκτονική του μετατροπέα. Μελέτες που έχουν γίνει δείχνουν ότι ο θόρυβος υπό προϋποθέσεις εξαρτάται από το μέτρο της εισόδου και όταν αυτή είναι ακέραια υποδιαίρεση του εύρους του κβαντιστή τότε ο θόρυβος έχει μικρή ισχύ, ενώ όταν είναι πολύ κοντά σε ακεραία υποδιαίρεση έχει σημαντικά μεγαλύτερη. Το παραπάνω ισχύει για αργά μεταβαλλόμενα σήματα, δηλ., όταν η συχνότητα δειγματοληψίας ξεπερνάει κατά πολύ τη συχνότητα του σήματος.

Για την μετατροπή της ψηφιακής εξόδου του κβαντιστή σε αναλογική μορφή χρησιμοποιούνται βαθυπερατά φίλτρα καταλλήλου εύρους, έτσι ώστε να αποκόπτουν τις απότομες μεταβολές της εξόδου του κβαντιστή και να διατηρούν τη χρίσημη πληροφορία που εμπεριέχεται σε αυτή. Συχνά χρησιμοποιείται και ένα ενδιάμεσο στάδιο που ομαδοποιεί σε λέξεις ένα πλήθος από bits, μειώνοντας έτσι τη συχνότητα της εξόδου. Με αυτό τον τρόπο χαλαρώνονται οι απαιτήσεις για τη σχεδίαση του βαθυπερατού φίλτρου που ακολουθεί, και επιπρόσθετα υπό προϋποθέσεις μπορούμε να πετύχουμε μια μικρή μείωση του θορύβου που υπάρχει σε συχνότητες κοντινές με αυτή του σήματος. Ιδιαίτερη προσοχή θα πρέπει να δοθεί σε αυτό το στάδιο, καθώς αν δεν λάβουμε υπόψη ορισμένες παραμέτρους, ο θόρυβος ενισχύεται.

Ιδιαίτερη σημασία επιπλέον πρέπει να δοθεί και στην ανάλυση ευστάθειας. Να σημειώσουμε ότι εφόσον ο βρόχος εμπεριέχει μη γραμμικό στοιχειό (κβαντιστής), το σύστημα δεν

μπορεί να μοντελοποιηθεί ως γραμμικό, και δεν μπορούν να χρησιμοποιηθούν γνωστά κριτήρια ευστάθειας. Αυτό που μας ενδιαφέρει είναι, μέχρι ποια τιμή μπορεί να φτάσει το μέτρο της εισόδου σε σχέση με το εύρος του κβαντιστή, έτσι ώστε να εξασφαλίζεται ότι οι εσωτερικές καταστάσεις του μετατροπέα θα παραμείνουν σε ορισμένο εύρος τιμών. Όταν ο μετατροπέας φτάνει σε αστάθεια, οι καταστάσεις του οδηγούνται στο άπειρο και η έξοδος του κβαντιστή μένει σταθερή είτε στην μέγιστη είτε στην ελάχιστή της τιμή. Να σημειώσουμε ότι δεν υπάρχουν ξεκάθαρα όρια για το που υπερβαίνεται η ευστάθεια, καθώς κοντά στα όρια ευστάθειας-αστάθειας υπάρχουν κάποιες τιμές για τις οποίες η ευστάθεια εμφανίζεται ορισμένες φορές και μετά από 20.000 περιόδους προσομοίωσης, για τον μετατροπέα τρίτου βαθμού με κβαντιστή 15 επιπέδων που χρησιμοποιήθηκε. Να σημειώσουμε ότι οι προσομοιώσεις σε τέτοια κλίμακα είναι επηρεπείς σε αριθμητικά σφάλματα που εισάγονται λόγω της αριθμητικής πεπερασμένης ακριβείας της υπολογιστικής μονάδας. Έτσι είναι αναγκαίο να εισάγουμε επιπρόσθετο θόρυβο, που υπάρχει και στην πραγματικότητα στα ηλεκτρονικά κυκλώματα προκειμένου να ξεπεραστεί αυτή η δυσκολία.

ΠΑΡΑΡΤΗΜΑ

Table of Contents

Introduction

In digital photography, the light's intensity level is stored in small sensors, referred as pixels. For a more "close to reality" conversion, apart from the total number of these pixels, we are also interested for the bit-accuracy that each pixel can store. Latest digital cameras maximum resolution is 22 bits, achieved by oversampling analog to digital converters [16].

Today, one could say that digital processing has become the cheapest and most efficient way to process all types of signals. Since most signals are analog in nature, we require an analog to digital stage, in order to convert these signals in digital form. So, if we have a B-bit analog to digital converter (ADC), we can represent the theoretically infinite values in analog range, to 2^B values in digital form.

Several architectures of ADC have been developed, depending on the requirements of each application. These requirements may concern the input signal characteristics like amplitude and frequency. They may also concern accuracy in bits of the resulting digital signal.

There are two main categories of ADC, with respect to the number of samples taken. First, those that sample the analogue signal with at least twice the frequency of signal's highest frequency, i.e., at least two samples per period as extracted by Nyquist sampling theorem. These ADC have a transfer function as shown in Fig. 1.4.

Each value of the analog input signal is quantized in one of the converter's levels (2^B in number). For every added bit, hardware increases logarithmically. This is not a problem by itself, but if we take into acount the element matching required for big cirquits, we can understand the difficulty of these architectures.

The second category are converters sample at a frequency n times higher than Nyquist rate. These are called oversampling modulators, and n is called the oversampling ratio (OSR). These modulators use 2 OSR samples per period, and use this abundance of samples

in order to achieve higher accuracy, usually more than 20 bits. These modulators use converters of a few bits in accuracy in feedback loops, and are studied in depth in this thesis.

State-of-the art ADC can reach 20 Gs/s for ENOB = 6.5 bits [6], at the expense of power consumption that reaches 10 W. Other designs with less than 1 W power cinsumption, reach 1 Gs/s for ENOB = 8.85 bits [7], or 0.8 Gs/s for ENOB = 9 bits [5], or 1.35 Gs/s for ENOB = 7.7 bits [9], or by imroving the last design 1.8 Gs/s for ENOB = 8.3 bits [15]. On the other hand, in the area of high resolution ADC, we find architectures that have ENOB = 20 or 22 bits, at the expense of slow input data rate, e.g., 15 Hz [11] or 12.5 Hz [10]. More balanced architectures, described in [12], can have ENOB around 15 bits for input frequency up to 1 MHz.

Concerning our work, in the first chapter, based on [13], we present the basic operation of a simple oversampling modulator, including operational equations and basic accuracy and noise approaches. In the second chapter we study in more detail the quantization noise, first using general noise analysis and then taking into account system dynamics that affect noise statistics. In the third chapter we examine a specific type of decimation filters, called $sinc^k$. along with their effect in baseband noise. Finally we study stability issues, concerning sufficient conditions that keep the system stable.

Due to the lack of general theoretical results concerning some aspects of the operation of oversampling ADC, our work is based on extensive simulation analysis, verifying whatever studied and presented in this thesis.

Chapter 1

Oversampling

1.1 Basic Idea

Oversampling ratio (OSR) is how many times above the Nyquist rate, we sample an analog signal in order to achieve analog to digital conversion (ADC). This aims to achieve higher resolution in bits, keeping circuit complexity low at the expense of higher sampling rate. So oversampling is ideal for low-frequency signals, like audio ones, since a 100 MHz signal with OSR = 64 would require $64 \cdot 2 \cdot 100\,\text{MHz} = 12.8\,\text{GHz}$ sampling rate which is very difficult to achieve.

1.2 Single stage delta-sigma modulator

Delta-Sigma modulators are the basic type of oversampling converters used today, and they have the form of Fig. 1.1.

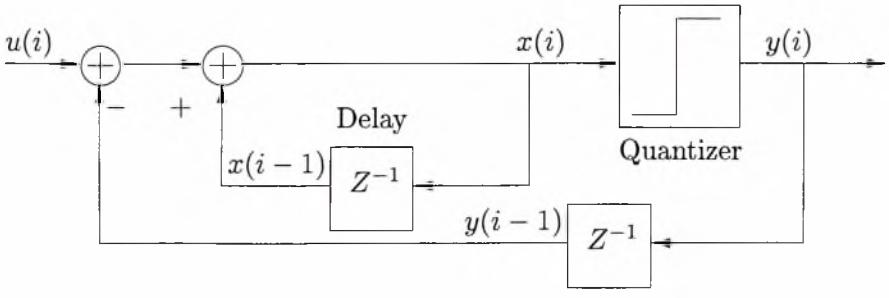

Figure 1.1: Single stage oversampling ADC

Input $u(i)$ is a discrete time signal. The inner loop, consisting of the delay element along with positive feedback, can be considered as an integrator in digital form, since it adds the previously created state $x(i-1)$ to the new input. For further details concerning continuous and discrete time conversions refer to [14]. Furthermore consider the quantizer as a comparator deciding $+1\,\mathrm{V}$ if its input (i.e., $x(i)$) is positive or zero and $0\,\mathrm{V}$ if negative. The functional equations describing the modulator in Fig. 1.1 are

$$y(i) = Q[x(i)] \tag{1.2.1}$$

$$x(i) = x(i-1) + u(i) - y(i-1) \tag{1.2.2}$$

Note that since the quantization function $Q[x(i)]$ is inherently non-linear, behavior of the circuit can't be predicted with standard state-space techniques.

1.3 Behavior of delta-sigma modulator for steady input

Assuming $u(i) = 0.25\,\mathrm{V}$ and $x(1) = 0$, for $i = 1:4$, the functional equations give:
$y(1) = Q[x(1)] = Q[0] = 1,$
$x(2) = x(1) + u(2) - y(1) = 0 + 0.25 - 1 = -0.75$
$y(2) = Q[x(2)] = Q[-0.75] = 0,$
$x(3) = x(2) + u(3) - y(2) = -0.75 + 0.25 - 0 = -0.5$
$y(3) = Q[x(3)] = Q[-0.5] = 0,$
$x(4) = x(3) + u(4) - y(3) = -0.5 + 0.25 - 0 = -0.25$
$y(4) = Q[x(4)] = Q[-0.25] = 0,$
$x(5) = x(4) + u(5) - y(4) = -0.25 + 0.25 - 0 = 0$

The resulting output of the delta-sigma loop for four samples of the input is $y(i) = \{1, 0, 0, 0\}$, having a mean value $1/4 = 0.25\,\mathrm{V}$, which perfectly matches the input. Note that since $x(5) = x(1)$ the above pattern will continue unchanged for future samples. In Fig. 1.2 we have started to plot from sample 64 in order to avoid seeing transient effects at the output of the filter. The filter is a first order low-pass following the quantizer output. Note that the oscillation of the filter depends on its cuttof frequency. Since the input signal is constant, the normalized cuttof frequency is expressed as $1/2\mathrm{OSR}$ (the bigger the OSR

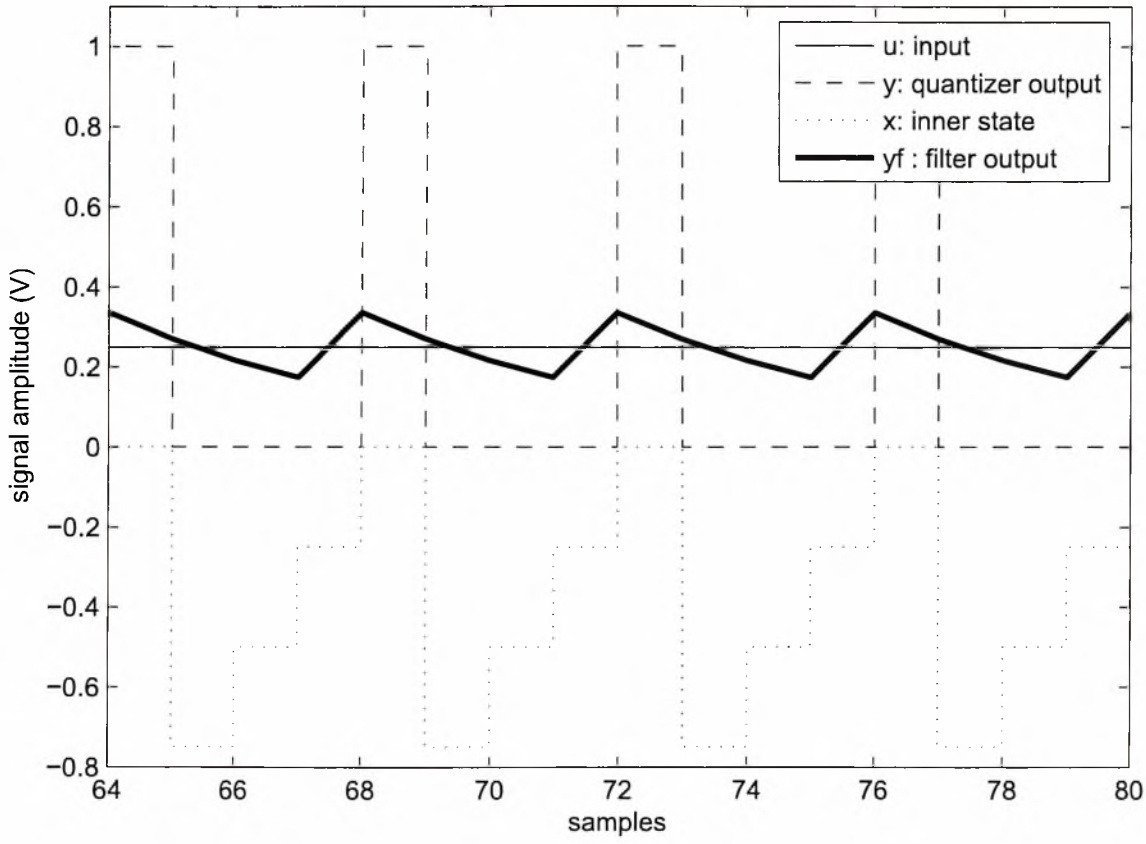

Figure 1.2: Quantizer and filter output for u=0.25V

the narrower the filter). In Fig. 1.2 we used OSR = 64 and from oscillation's amplitude we have calculated the signal-to-noise ratio: SNR = 41.2 dB and resolution: 12.3 bits. For different values of OSR, simulation results are shown in Table 1.1. It is clear that higher OSR leads to better accuracy. Also note that using a single bit quantizer we have reached acuracy of 18 bits.

If $u(i) = 1.5V$ and $x(1) = 0$, from the functional equations we have:

$y(1) = Q[x(1)] = Q[0] = 1,$

$x(2) = x(1) + u(2) - y(1) = 0 + 1.5 - 1 = 0.5$

$y(2) = Q[x(2)] = Q[0.5] = 1,$

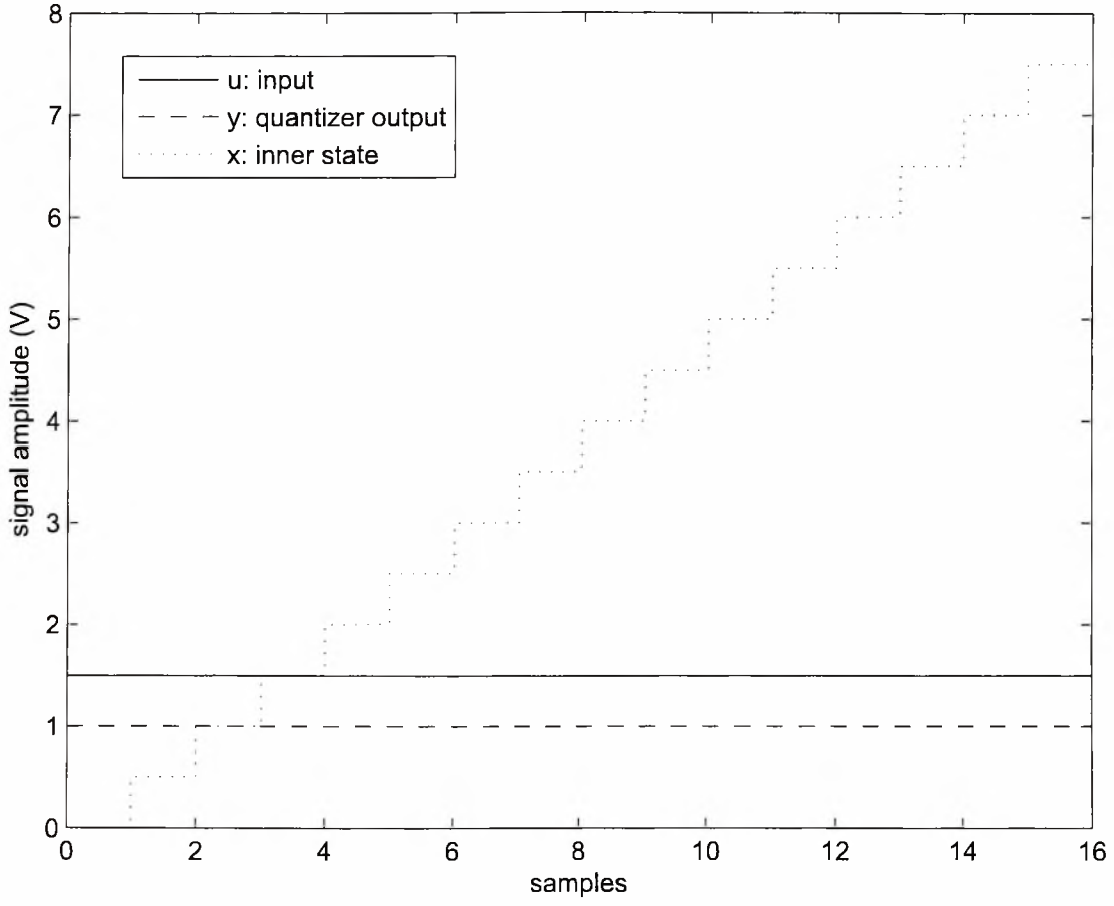

Figure 1.3: Quantizer and filter output for u=1.5V

$x(3) = x(2) + u(3) - y(2) = 0.5 + 1.5 - 1 = 1$
$y(3) = Q[x(3)] = Q[1] = 1,$
$x(4) = x(3) + u(4) - y(3) = 1 + 1.5 - 1 = -1.5$

As plotted in Fig. 1.3 we see that state x tends to infinity and output y is always 1. We say that input $u(i) = 1.5V$ is leading the modulator to instability. Stability for this simple type of modulator is assured for any input in the range of $0 \leq u(i) \leq 1$ since $u(i) = 0$ would give all-zero and $u(i) = 1$ all-one output.

OSR	SNR	bits
4	17.8	5.9
16	29.2	9.7
64	41.2	13.6
256	53.2	17.7
512	59.3	19.7

Table 1.1: Simulation results for SNR in dB and resolution in bits for steady input

1.4 B-bit symmetric quantizer

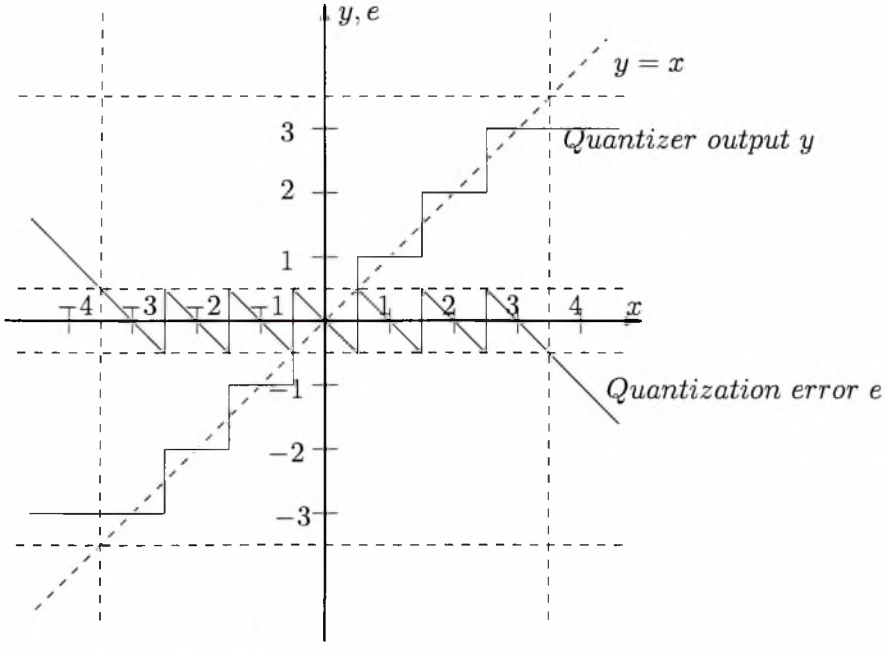

Figure 1.4: 3-bit symmetric quantizer

Fig. 1.4 shows the transfer function of a B-bit symmetric quantizer with B =3. Straight line $y = x$ represents perfect matching of the input x to the output y, but since the quantizer can only take seven values, quantization error e occurs, which is the difference between the ideal output ($y = x$) and the the actual output of the quantizer. The difference between output thresholds is called least-significant bit (LSB) size and in this case is $\Delta = 1$. The maximum level of the quantizer output is $M = 2^{B-1} - 1 = 3$ and the minimum is $-M$ respectively. Note the quantization error is between $-\Delta/2$ and $\Delta/2$ when input stays within

the range $-M - \Delta/2 \leq x \leq M + \Delta/2$.

We introduced this type of of quantizer in order to show significant differences in low-frequency noise shape after the filtering process. In Fig. 1.5 and Fig. 1.6 we see a full period of a sinewave with 256 samples per period. OSR in both these cases is 32, and the signal is considered to be 4 times slower than maximum input frequency. In Fig. 1.5 we used a simple 1-Bit quantizer to sample a sinewave of 0.5 Vpp amplitude and a 0.5 V dc offset in order to stay within the stability margins. Using a first order low-pass filter with cuttof frequency coresponding to OSR = 32, we obtain SNR = 20 dB and 9.5 bits of accuracy. For OSR = 64 we obtain SNR = 24 dB and 10.8 bits. In Fig. 1.6 we used the 3-Bit quantizer of Fig. 1.4 to sample a centered 5 Vpp amplitude sinewave. Same as before, for OSR = 32, SNR = 28 dB and 8.5 bits, while for OSR = 64, SNR = 29 dB and 9.1 bits.

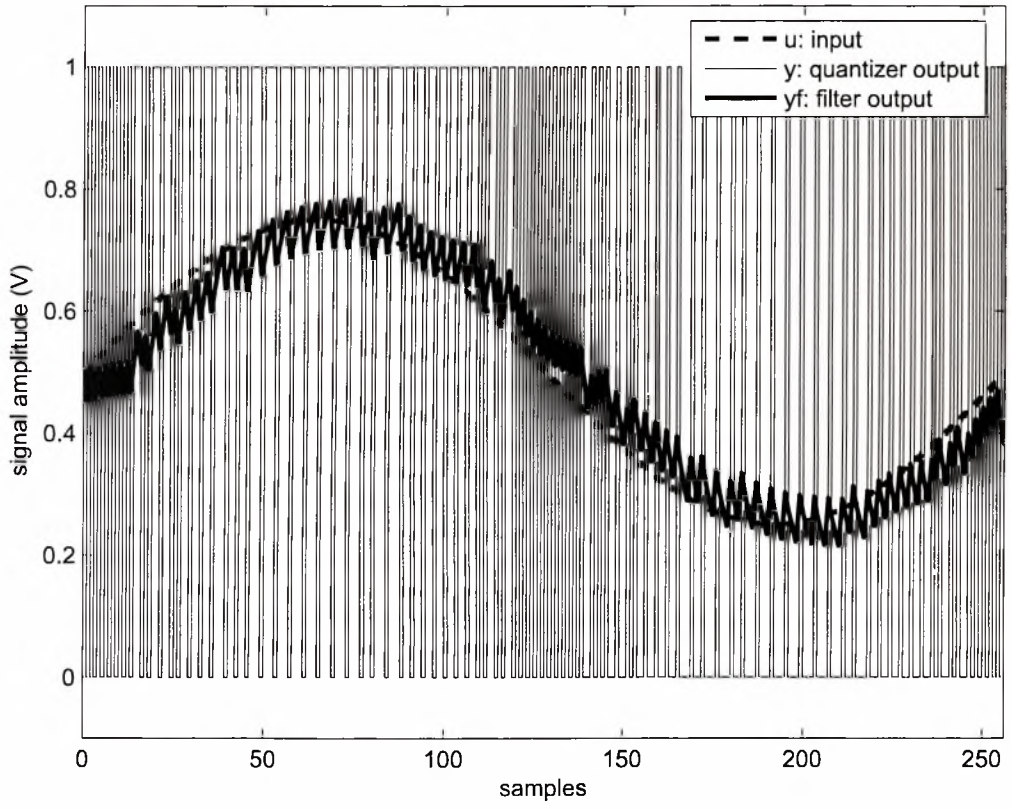

Figure 1.5: Oversampling a sinwave with a 1-Bit quantizer

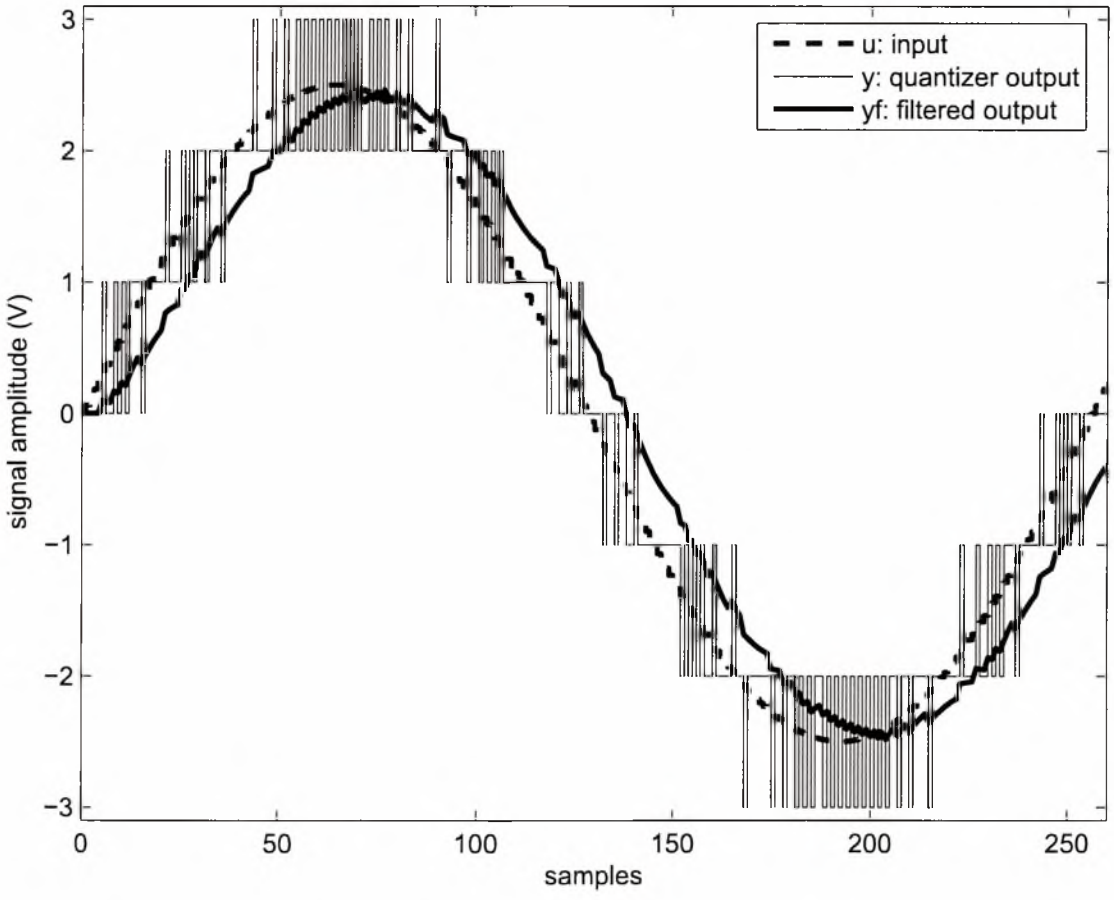

Figure 1.6: Oversampling a sinwave with a 3-Bit quantizer

Chapter 2

Noise

2.1 Basic considerations

As noted in the previous chapter, we are highly interested in the quantization noise that lies within the passband of the low-pass filter. Its magnitude determines the signal to noise ratio (SNR), or equivantly the accuracy in bits that we finally achieve. At first we will study the baseband quantization noise for a specific type of modulator and later on we will see that there are factors, that when taken into consideration, affect our basic assumptions.

2.2 Noise in first order modulator

For the 3-bit symetric quantizer shown in Fig. 1.4 of the previous chapter, we saw that quantization error is between $-\Delta/2$ and $\Delta/2$ when input stays within the range $-M-\Delta/2 \leq x \leq M+\Delta/2$. Since x can take randomly any value, depending also on the feedback gain as shown in [1], e can be considered as having zero mean and mean-square $\sigma_e^2 = \Delta/12$. Using this linear model shown in Fig. 2.1 for the quantizer, the first order modulator is described in z-domain by following equation

$$\begin{aligned} Y(z) &= X(z) + E(z) = z^{-1}X(z) + U(z) - z^{-1}Y(z) + E(z) \\ &= U(z) - z^{-1}[Y(z) - X(z)] + E(z) = U(z) + (1 - z^{-1})E(z) \end{aligned} \tag{2.2.1}$$

Compared to the general form: $Y(z) = \mathrm{STF}(z)U(z) + \mathrm{NTF}(z)E(z)$, the signal transfer function is $\mathrm{STF} = 1$ and the noise transfer function if $\mathrm{NTF} = 1 - z^{-1}$, which is a high-pass response having $|NTF|^2 = (2\pi f)^2$ for $z = e^{j2\pi f}$ and $f \ll 1$.

For the linear model of quantizer, the spectral density of e is $S_e(f) = 2\sigma_e^2 = 2/3$ for

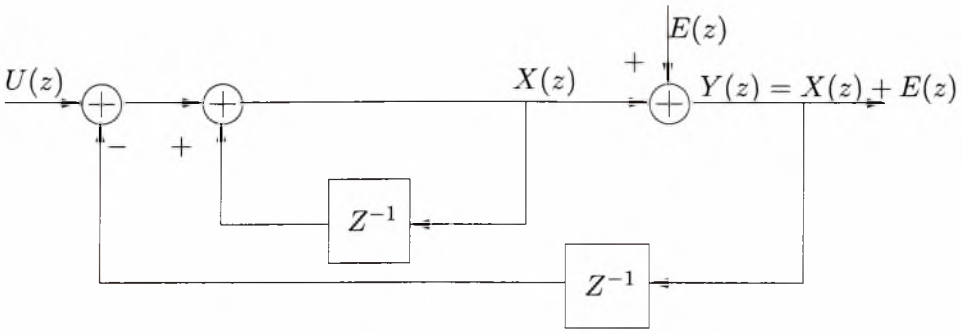

Figure 2.1: First order modulator linear model

$\Delta = 2$. If the cut-off frequency of the low pass filter is $f_c = 1/(2 \cdot \text{OSR})$, the in-band noise power is

$$\sigma_q^2 = \int_0^{f_c} (2\pi f)^2 S_e(f) df = \frac{\pi^2}{9(\text{OSR})^3} \tag{2.2.2}$$

Consider a sinewave input with peak amplitude M. Since STF $= 1$, its power is $\sigma_u^2 = M^2/2$, therefore

$$\text{SNR} = \frac{\sigma_u^2}{\sigma_q^2} = \frac{9M^2(\text{OSR})^3}{2\pi^2} \tag{2.2.3}$$

Note that the accuracy can be expressed as the effective noumber of bits (ENOB). The relationship between SNR and ENOB for sinewave excitation is SNR $= 6.02\text{ENOB} + 1.76$ [13, sec. 1.1]. So, for doubling the OSR, SNR is increased by 9 dB and ENOB by 1.5 bits.

2.3 Noise in second order modulator

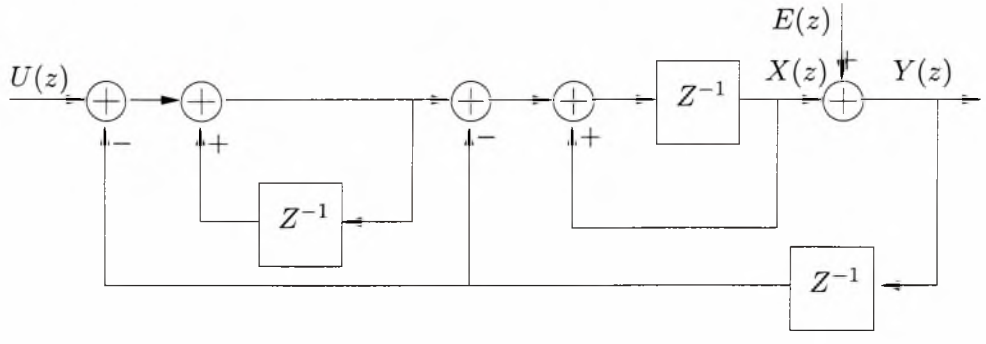

Figure 2.2: Second order modulator linear model

For the second order modulator shown in Fig. 2.2, the output equation is: $Y(z) =$

$z^{-1}U(z) + (1 - z^{-1})^2 E(z)$. It is clear that there is a one sample delay from the input u to the output y since there is a delay element on the straight-forward path. Note also that since NTF $= (1 - z^{-1})^2$ has two zeros at dc (instead of one in the first order modulator), we expect increased attenuation of quantization noise at low frequencies. Same as before $|\text{NTF}|^2 = (2\pi f)^4$ for $f \ll 1$, and

$$\sigma_q^2 = \int_0^{f_c} (2\pi f)^4 S_e(f) df = \frac{2\pi^4}{15(\text{OSR})^5} \tag{2.3.1}$$

For a sinewave input with peak amplitude M

$$\text{SNR} = \frac{M^2/2}{\sigma_q^2} = \frac{15M^2(\text{OSR})^5}{4\pi^2} \tag{2.3.2}$$

The ENOB can be found as in the first order modulator case. Note that in the second order case, for doubling the OSR, SNR is increased by 15 dB and ENOB by 2.5 bits. In general, the higher the order of the modultor the greater the increase in SNR with OSR. We should keep in mind that when adding feedback loops, bigger stability constraints arise for input signal and quantizer structure.

2.4 Experimental results

For the two modulators described in previous section, Fig. 2.6 shows experimental results concerning quantization noise. Since the first order modulator has NTF with a single zero at dc and the second order has NTF with a double zero at dc, we except their diagram to have 20 dB/dec inclination and 40 dB/dec respectively. On the right diagrams we have modeled the quantizer as a source of zero-mean noise with $\sigma_e^2 = \Delta/12$ where $\Delta = 2$. Note that for the simulation we have used a sinewave of the form $u = 0.9\sin(T_p 2\pi t/T)$, having $T_p = 100$ periods of simulation, and $T = 2(2 + 3/17) \cdot T_p \cdot \text{OSR}$ total number of samples, with OSR $= 128$. Notice that we used a bit more than $4 \cdot \text{OSR}$ samples per period, because this would result in misleading results on the FFT analysis. On the left diagrams we didn't use any model for the quantization noise and just simulated the circuit with a single-bit quantizer using the *sign* operation. This is why $\Delta = 2$.

It is obvious that there is some harmonic distortion higher than the noise level. This shows that quantization noise isn't totaly random as previously considered. The above analysis though, shows how the relation between OSR and σ_q^2 changes for different modulator architectures.

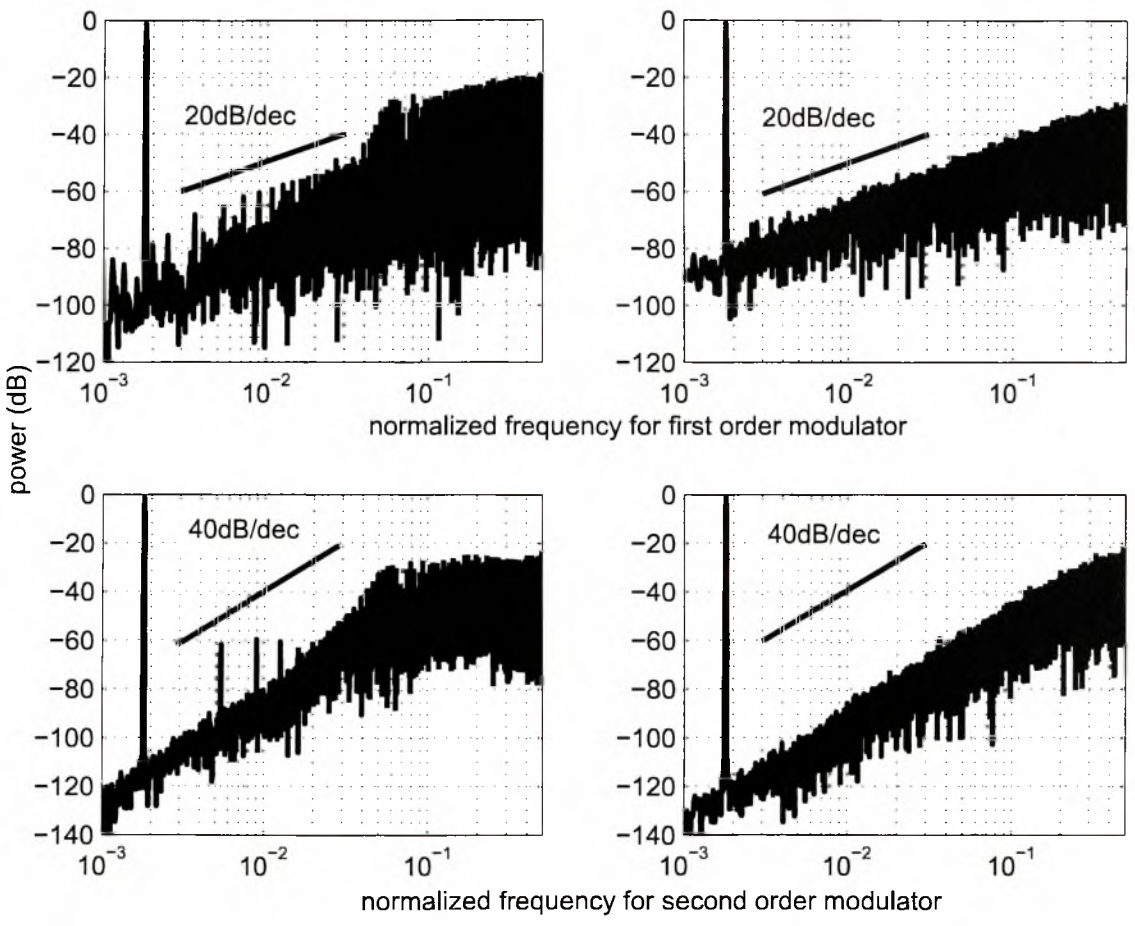

Figure 2.3: simulation results

2.5 Quantization noise for slow-changing signals

As proved in [4], for steady input, the quantization noise is highly correlated with the amplitude of the input signal. The same result also holds when the sampling rate of the oversampling modulator far exceeds the frequency of the input signal. Additional noise analysis can be found in [2].

Consider the modulator shown in Fig. 2.4 ignoring the clocking of impulse generator. Since u is a steady input of magnitude u, the output x of the analog integrator would be a ramp, as shown in Fig. 2.5. Whenever signal x becomes positive, the impulse generator

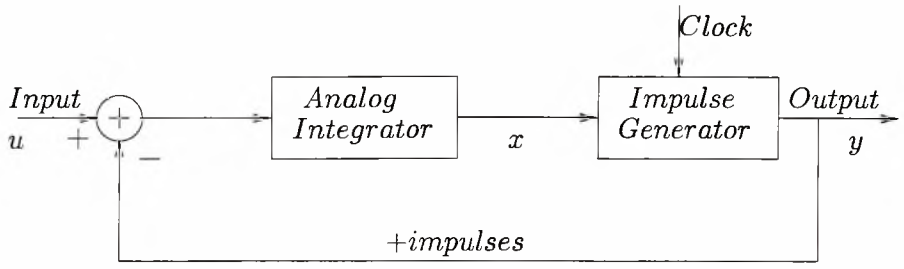

Figure 2.4: First order integrating modulator

creates an impulse y of magnitude A, which is subtracted from input u. Increasing the input u, leads to steeper output x of the integrator and thus impulses occur more often. Impulse frequency is x/A.

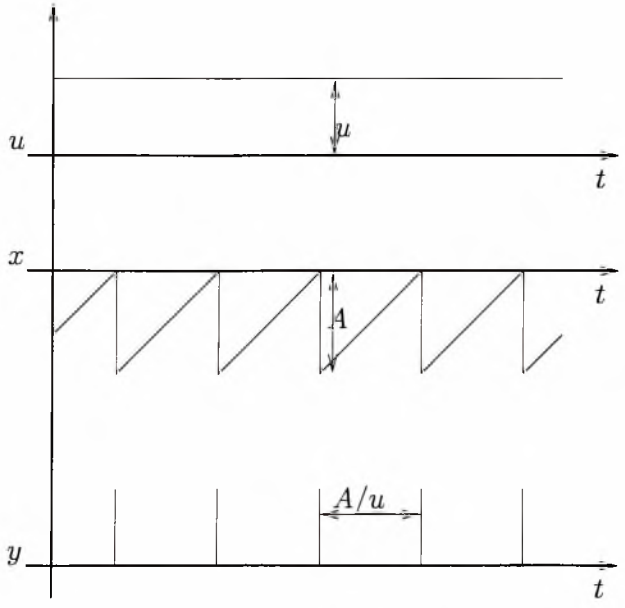

Figure 2.5: modulator in continuous time

If we include clocking of period τ, as in Fig. 2.6, an impulse is generated only when the clock is present. If u is exactly A/τ, an impulse occurs in each clock instance. If it is slightly more, an impulse would still occur in each clock, but in each period we have a positive remainder summing up to infinity. Note that impulses still occurs at an average rate of x/A. The same average rate, can be generated by sampling a sequence of rectangular pulses R of frequency x/A and duration τ as shown in Fig. 2.7. This is also proved by

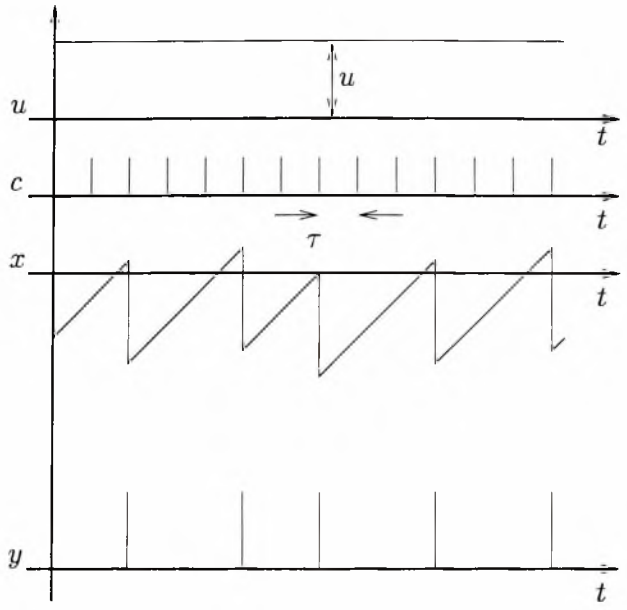

Figure 2.6: modulator in discrete time

simulation results.

We use this equivalence (that is proved later on also by simulation results), in order to derive an easier way to describe modulation noise with mathematical equations.

The output y can be expressed as the product of $C(t)$ and $R(t)$ from Fig. 2.7. Ignoring constant delays, y can be expressed as

$$\begin{aligned} y(t) = C(t)R(t) &= \sum_k \exp(\frac{2\pi jkt}{\tau}) \sum_l \frac{\sin(\pi lu)}{\pi l} \exp(\frac{2\pi jlut}{\tau}) \\ &= \sum_l \sum_k \frac{\sin(\pi lu)}{\pi l} \exp(2\pi j \frac{lu+k}{\tau} t) \end{aligned} \tag{2.5.1}$$

The result in Eq. 2.5.1 represents the output signal as a set of spectral lines of frequency $f = (lu + k)/\tau$ and since we are interested in the band of frequency that is less or equal than half the sampling rate, i.e., $f \leq 1/2\tau$, y can be expressed as

$$y(u) = u + 2 \sum_{l=1} \frac{\sin(\pi lu)}{\pi l} \cos(2\pi [lu] \frac{t}{\tau}) \tag{2.5.2}$$

where the first term is the useful output and the second is modulation noise. Note that $[u]$ represents the fractional roundoff of a real number u having nearest integer $I(u)$. For further details refer to [4].

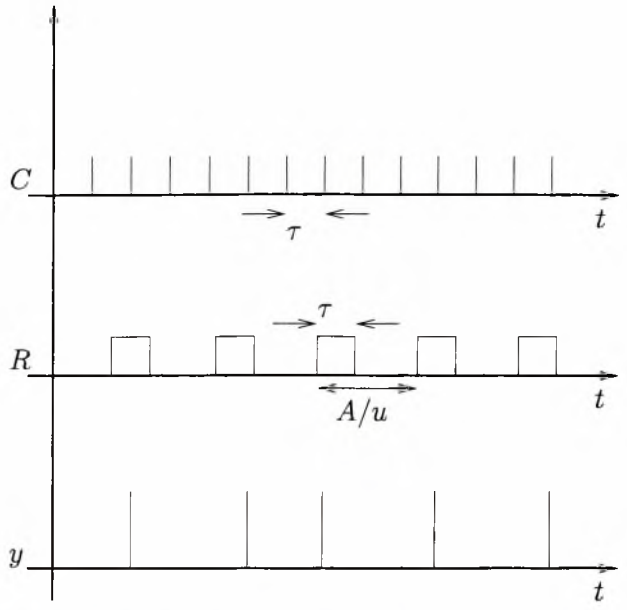

Figure 2.7: Equivalent to sampling at discrete time

Let f_0 be the baseband, i.e., the cutoff frequency of a low-pass filter following the modulator. For the noise component of y in Eq. 2.5.1 of frequency $f = [lu]/\tau$ to lie in the baseband f_0, it is required that

$$|[lu]| < f_0\tau \tag{2.5.3}$$

and the component's associated power is

$$P_l(lu) = 2\frac{\sin^2(\pi[lu])}{(\pi l)^2} \tag{2.5.4}$$

Note that l is an integer and the larger its value, the smaller the power component given by Eq. 2.5.4. Since the input range is $0 \leq u \leq 1$, we want to find those integers l for which, for a given value of the input, inequality 2.5.3 is satisfied. Suppose $f_0\tau$ is 0.1 and u is 0.5. Inequality 2.5.3 is satisfied for even values of l for which $[lu] = 0$ resulting in zero noise components. If u is slightly bigger or smaller than 0.5 then non zero noise components are present for even values of l, that have to be summed up in order to find the noise power for the specific input amplitude.

In Fig. 2.8 we have plotted noise power in dB according to equations 2.5.4 and 2.5.4 for the whole input range, for a baseband $f_0 = 3.5\,\text{KHz}$ and for sampling rates of 64, 256 and

512 KHz. For each case τ is the inverse of the sampling rate, resulting in smaller $f_0\tau$, which is the bound of inequality Eq. 2.5.3.

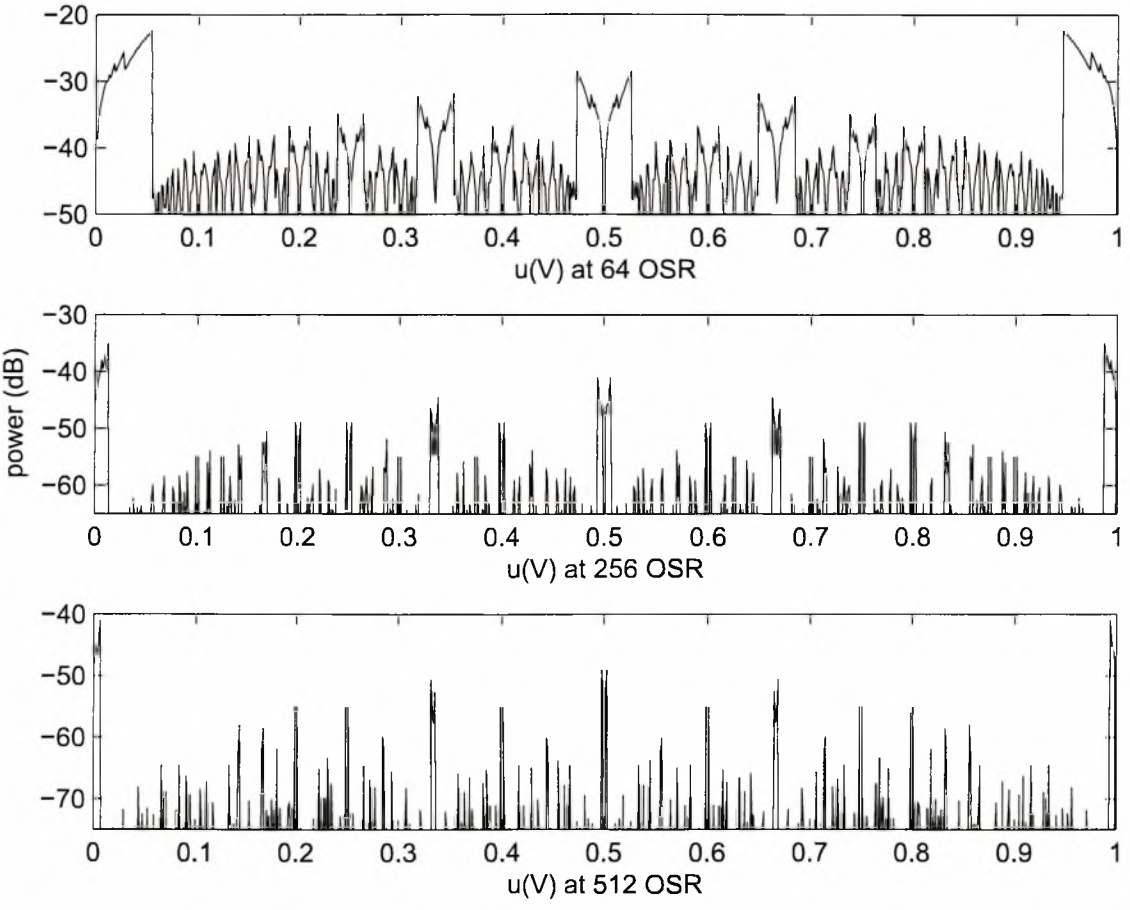

Figure 2.8: Results from theoretical Eq. 2.5.4 for OSR = 64, 256 and 512

Observations concerning the correlation between baseband noise and input amplitude reveal that noise power is close to zero in integer divisions of the input signal and has peak values in the vicinity of these divisions. As $f_0\tau$ becomes smaller, i.e., the sampling rate increases, these vicinities become narrower and peak values larger.

Next we run a simulation on the first order modulator using the *sign* quantizer for a full scale sinewave and OSR = 256. In Fig. 2.9 by comparing the two plots, we verify our assumptions.

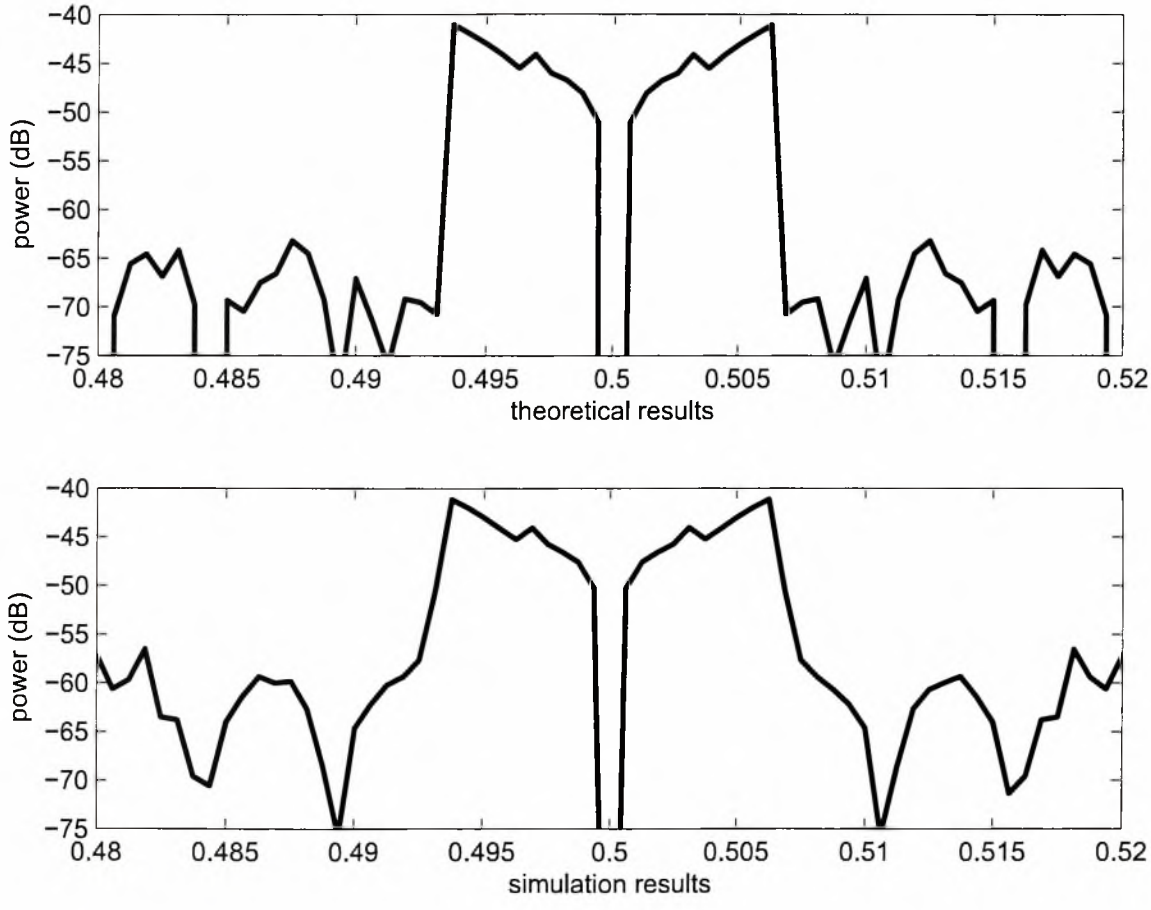

Figure 2.9: Comparison of theoretical and simulation results

Chapter 3

Decimation

3.1 Introduction

Decimation is the the trading of word length for word rate. Suppose we have a four bit quantizer sampling a signal of 1 MHz with oversampling ratio 512. The resulting signal is a four bit word length having 512 MHz word rate. This signal is quite difficult for further processing because of its high word rate. The most simple type of decimation is to take the mean value of every 64 samples and use these values instead of the original samples. This would reduce the the word rate to 8 MHz, but we need also to increase the word length in order to conserve resolution. So we use decimation, that can be implemented easily in digital ways, to downsample the oversampled data to about four times the Nyquist rate, and then we use a more simple low-pass filter (than the one needed without the decimation stage) for further processing. The $sinc^k$ filters consist the basic type of decimation filters. First we will study the structure and transfer function of $sink^k$ filters and right afterwards their effect in baseband noise.

3.2 $sinc^k$ filters

For $k = 1$ we have the most simple $sink^k$ filter. It consists of $N - 1$ delays and computes the running average of the input data stream $y(n)$ which is the output of the quantizer for our case. The output $w(n)$ of the $sinc$ filter can be expressed as

$$w(n) = \frac{1}{N} \sum_{i=0}^{N-1} y(n-i) \tag{3.2.1}$$

having impulse response

$$h_1(n) = \begin{cases} 1/N, & \text{if } 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} \tag{3.2.2}$$

and z-domain transfer function

$$H_1(z) = \frac{1}{N} \frac{1 - z^{-N}}{1 - z^{-1}} \tag{3.2.3}$$

Its basic advantage is that it can be realized easily when used after a single-bit quantizer by just using discrete counters and registers. Fig. 3.1 shows an implementation of a sinc filter. The upper counter is incremented for each $+1$ from the quantizer and the lower counter, every N clock cycles, resets the upper and sends its value to the register. Thus, the output $w(n)$ of the register is N times slower than $y(n)$ and each of its value is an accounting of $+1$ produced by the quantizer during N sampling instances.

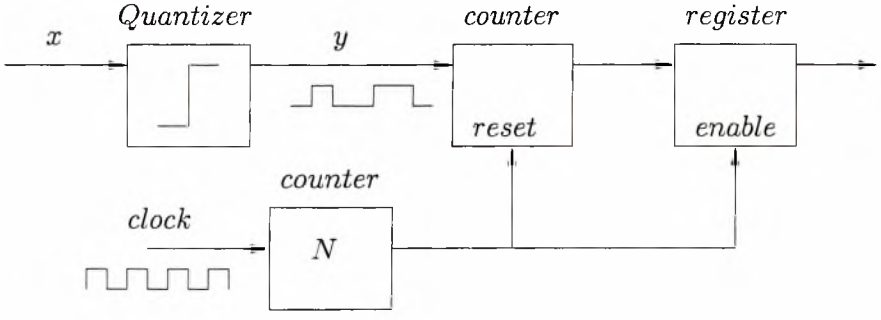

Figure 3.1: sinc filter implementation

For $k = 2$, $sinc^2$ is obtained by convolving the rectangular *sinc* filter by itself. The resulting triangular shape can be seen in Fig. 3.2. The z-domain transfer function is the one of the *sinc* filter squared

$$H_2(z) = \left(\frac{1}{N} \frac{1 - z^{-N}}{1 - z^{-1}}\right)^2 \tag{3.2.4}$$

Note that for $N = 16$, the *sinc* filter uses samples from $y(1)$ to $y(16)$ in order to form $w(1)$, $y(17)$ to $y(32)$ for $w(2)$ and so forth. On the other hand, in the $sinc^2$ filter, for the same $N = 16$, each sample is used twice since $w(1)$ needs samples $y(1)$ to $y(32)$, $w(2)$ needs samples $y(17)$ to $y(48)$, $w(3)$ needs samples $y(33)$ to $y(64)$, etc.

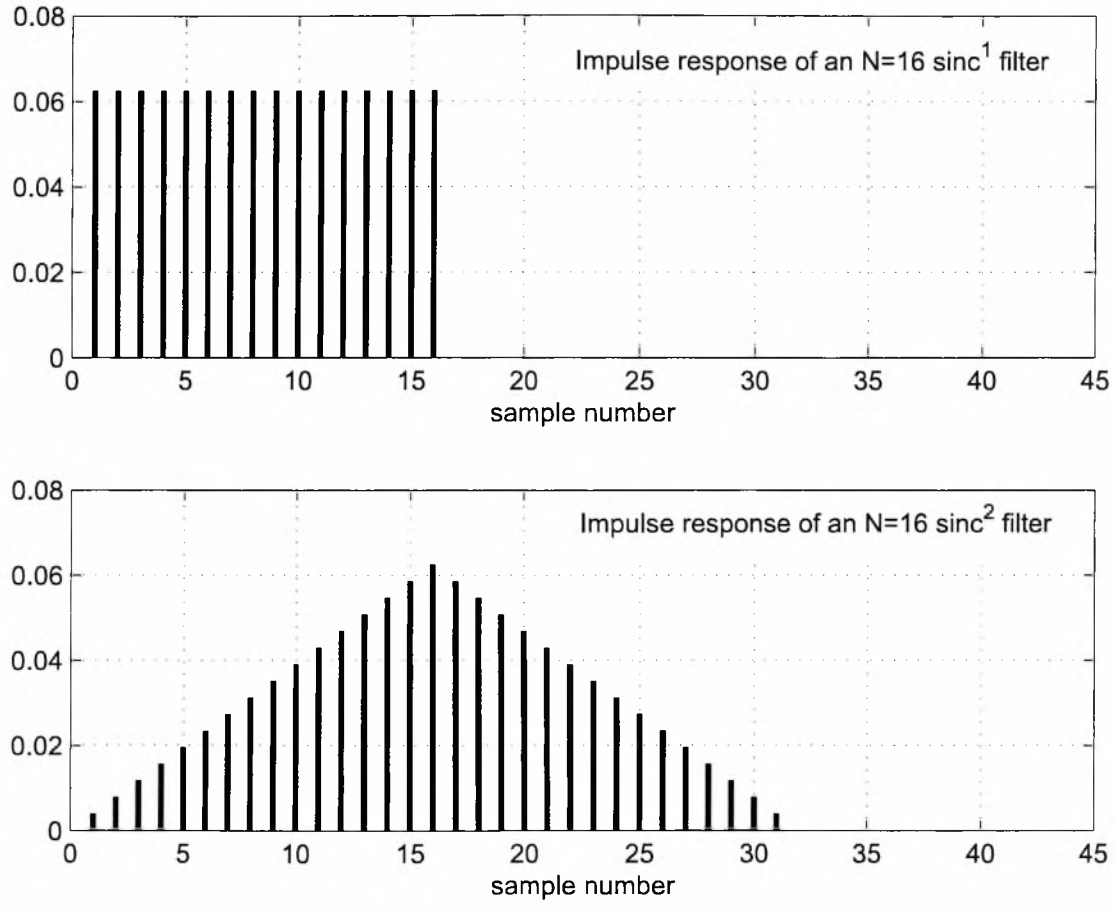

Figure 3.2: *sinc* and $sinc^2$ filters with N=16

For the higher order $sinc^k$ filters, we need k successive convolutions of the *sinc* filter, and their z-domain transfer function is

$$H_k(z) = (\frac{1}{N}\frac{1-z^{-N}}{1-z^{-1}})^k \tag{3.2.5}$$

3.3 Effect of $sinc^k$ filters on quantization noise

We need to judge on the effectiveness of a $sinc^k$ filter, in comparison with a low-pass filter, on a lth-order modulator. It follows from the noise analysis of the first and second order modulator, than for an lth-order modulator of the same architecture NTF $= (1-z^{-1})^l$ and

the noise at the output of the $sinc^k$ filter would be

$$Q_{kl}(z) = H_k(z)\mathrm{NTF}_l(z)E(z) = \frac{1}{N^k}\left(\frac{1-z^{-N}}{1-z^{-1}}\right)^k (1-z^{-1})^l E(z) \tag{3.3.1}$$

For $k = l$, Eq. 3.3.1 gives

$$Q_{l,l}(z) = \left(\frac{1-z^{-N}}{N}\right)^l E(z) \tag{3.3.2}$$

For $k = l+1$, Eq. 3.3.1 gives

$$Q_{l+1,l}(z) = \left(\frac{1-z^{-N}}{N}\right)^l \cdot \frac{1}{N}\frac{1-z^{-N}}{1-z^{-1}} E(z) \tag{3.3.3}$$

Notice that Eq. 3.3.3 differs from Eq. 3.3.2 in a factor equal to the transfer function of a *sinc* filter. So in $Q_{l+1,l}$ the noise component is averaged for every N samples. This is why the RMS value of $Q_{l+1,l}$ is $\sqrt{N}$ times smaller than the RMS of $Q_{l,l}$. These result are further analyzed in [3].

3.4 Experimental results

We used the same sinewave and modulators as in noise analysis, followed alternatively by either a *sinc* or a $sinc^2$ filter with $N = 17$. Fig. 3.3(a) shows the power spectral density of the output y of a first order modulator, while (b) and (c) shows the results of a *sinc* filter and $sinc^2$ respectively. In Fig. 3.4 we used the same filters at the output y of the second order modulator.

We should keep in mind that the basic role of this filter is not noise shaping, but to reduce the word rate. Since $N = 17$ we see that the filtered output is 17 times slower. For the second order modulator we see that for the *sinc* filter there is significantly more baseband noise. So as a rule of thumb, the order of the $sinc^k$ filter should be at least the same with the modulator's order.

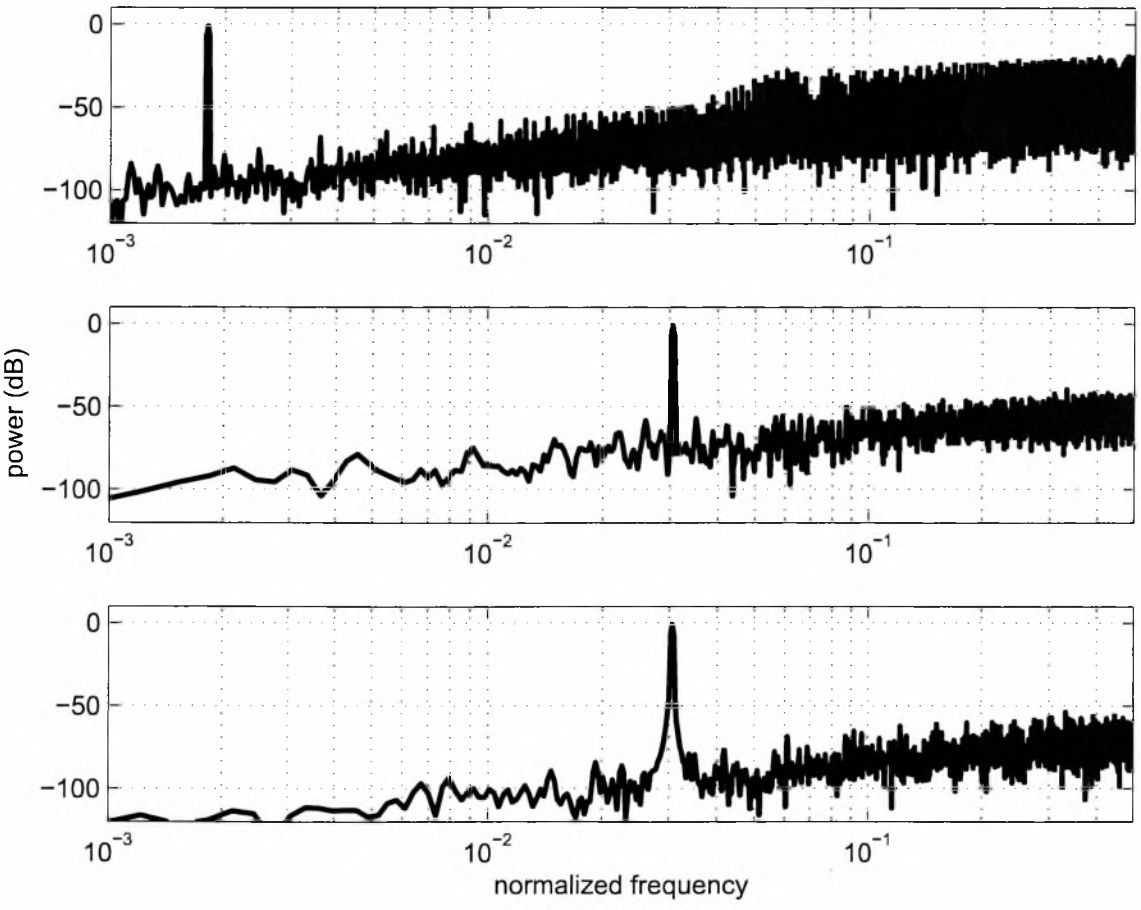

Figure 3.3: (a) first order modulator output, (b) effect of *sinc* filter , (c) effect $sinc^2$ filter

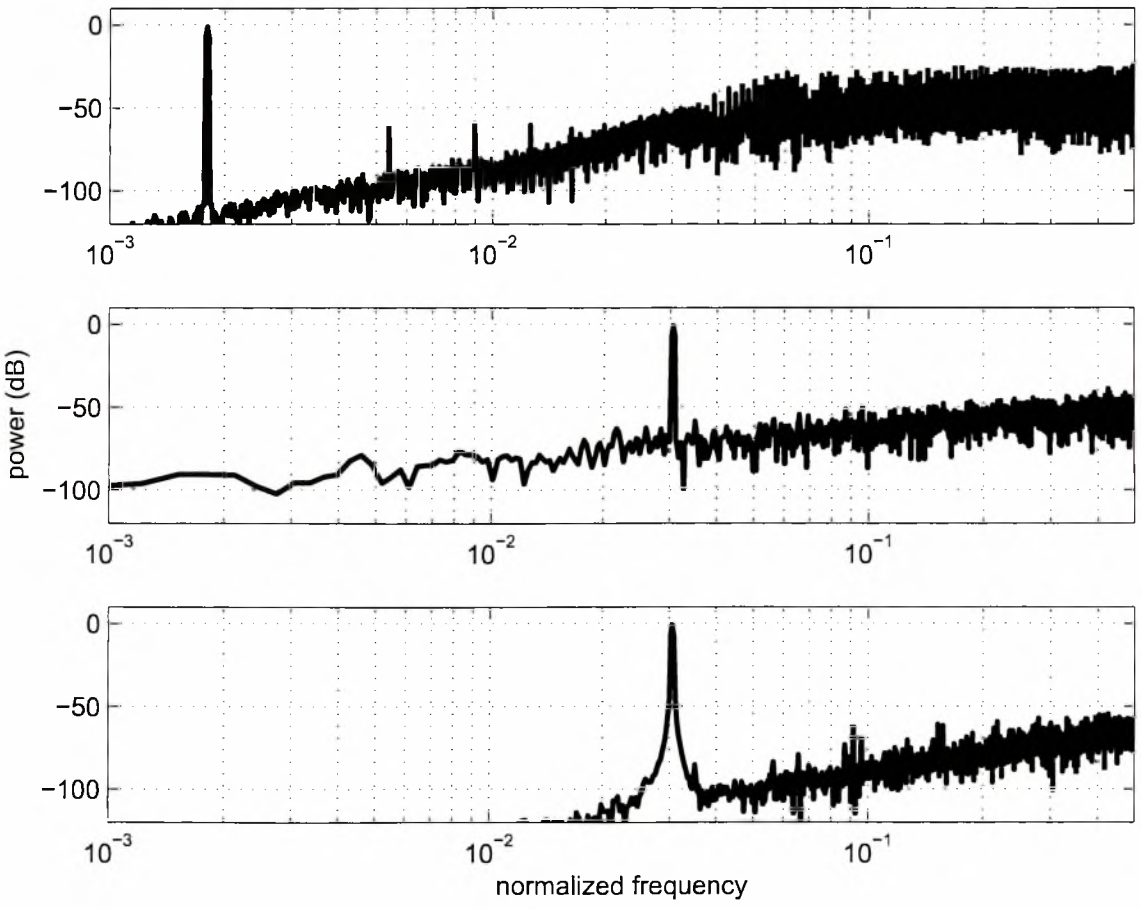

Figure 3.4: (a) first order modulator output, (b) effect of *sinc* filter , (c) effect $sinc^2$ filter

Chapter 4

Stability of oversampling converters

4.1 Theoretical analysis

As studied in [8] there exist sufficient conditions for which an oversampling converter of specific architecture remains stable. It can be proved that an N-th order modulator with NTF $= (1 - z^{-1})^N$, will remain stable if the quantizer has $B > N + 1$ bits and the input u is bound to half the quantizer no-overload input range.

The maximum input at quantizer is found to be limited by

$$\|x\|_\infty \leq \|STF\|_1 \cdot \|u\|_\infty + \|NTF\|_1 \cdot \|e\|_\infty - \|e\|_\infty \tag{4.1.1}$$

where STF is the signal transfer function, NTF is the noise transfer function, u the input signal and e the quantization error. Since quantizer input should not exceed the non-overload input range $|R|$, it is

$$|R| \leq \|STF\|_1 \cdot \|u\|_\infty + \|NTF\|_1 \cdot \|e\|_\infty - \|e\|_\infty \tag{4.1.2}$$

For the B-bit symmetric rounding quantizer we studied earlier, $|R| = (2^{B-1} - 1/2)$ and assuming that the input stays within the no-overload input range, the error is bound by $\|e\|_\infty = \Delta/2$.

For the modulator architecture of N stages using $N - 1$ delay-free integrators and one delaying just before the quantizer, the input-output relation is found to be

$$Y(z) = z^{-1}U(z) + (1 - z^{-1})^N E(z) \tag{4.1.3}$$

where

$$\mathrm{STF}(z) = z^{-1} \rightarrow \|\mathrm{STF}\|_1 = 1 \tag{4.1.4}$$

$$\mathrm{NTF}(z) = (1 - z^{-1})^N \rightarrow \|\mathrm{NTF}\|_1 = 2^N \tag{4.1.5}$$

Assuming the input signal u occupyes half the quantizer range, i.e., $\|u\|_\infty = 2^{B-2}$, inequality 4.1.2 gives

$$2^{B-1} - \frac{1}{2} \geq 1 \cdot 2^{B-2} + 2^N \cdot \frac{1}{2} - \frac{1}{2} \rightarrow B \geq N + 1 \tag{4.1.6}$$

We should note that the above condition for stability is sufficient but not necessary. After simulations on specific theoretical model, we check for how much can we exceed these boundaries and remain stable. Simulations near these boundaries may also give some insight concerning why, how and when an oversampling loop becomes unstable.

4.2 Simulation results

The above condition is verified by simulating a general N-th order modulator with a B-bit symmetric rounding quantizer for several values of $B > N + 1$ keeping $\|u\|_\infty = 2^{B-2}$.

Later on, we focused on a specific modulator, and tried several input signals. The modulator is a third order modulator, having two delay-free integrators and one delaying. The quantizer is a four bit symmetric rounding quantizer in order to satisfy the above condition. It has the form of Fig. 1.4 shown in an earlier chapter, with maximum level $M = 2^{B-1} - 1 = 7$ and minimum level $-M$, spaced with $\Delta = 1$. Starting with an input signal of magnitude $A = 2^{B-2} = 4$ (or 8 peak to peak), we progressively increased this value and notice when the system becomes unstable.

At first, for OSR = 128, we used a sinewave with relative frequency $f_n = 0.5$ meaning that it is two times slower than the maximum signal frequency. So this signal has exactly 512 samples per period. For $A \leq 5.46$ the system remained stable during 500 periods of simulation. For $A = 5.47$ it became unstable at the 295th period and for $A = 5.5$ it remained stable again. This fact is intuitively unjustifiable.

Firstly, we increased the time of simulation and noticed that for the same sinewave, instability starts around $A = 5.2$ after about 20.000 periods of simulation and still has stability regions mixed with instability ones around this region. The same happens with triangular and square signals, for any OSR.

Since in practice the input signal can't be perfectly synchronized with the sampling rate, and in order to avoid sampling at exactly the same point of the input signal in each period, we slightly changed the relative frequency of the signal f_n, so as not to have integer number of samples per period. Additionally, since any arithmetic inaccuracy is of the order of 1^{-10} or less, we add on the main stability loop some noise of magnitude 1^{-6}.

By using these adjustments, stability margins became more clear. There are no stability regions inside instability ones. Between the stability and the instability region, there is a zone of uncertainty leading sometimes to instability after 1.000 periods at the worst case. However, we still cannot derive any specific relationship connecting stability with signal magnitude A, relative frequency f_n and OSR. For OSR $= 128$ and $f_n = 0.497$ simulation results are summarized in table 4.2. The stability column has the maximum value of A, below of which the modulator remains stable. The instability column respectively, has the minimum value of A above which the modulator is unstable. Between these values there is this zone of uncertainty.

signal type	stability	instability
sinewave	5.20	5.60
triangular	5.26	5.80
rectangular	4.90	5.22

Table 4.1: Simulation results for stability and instability regions

In the diagrams shown in Fig. 4.1, we have plotted the maximum magnitude of the inner states (of the three inner states since we have a third order modulator). Notice that that the inner states are mainly influenced by the input and they are all positive since we have plotted absolute values.

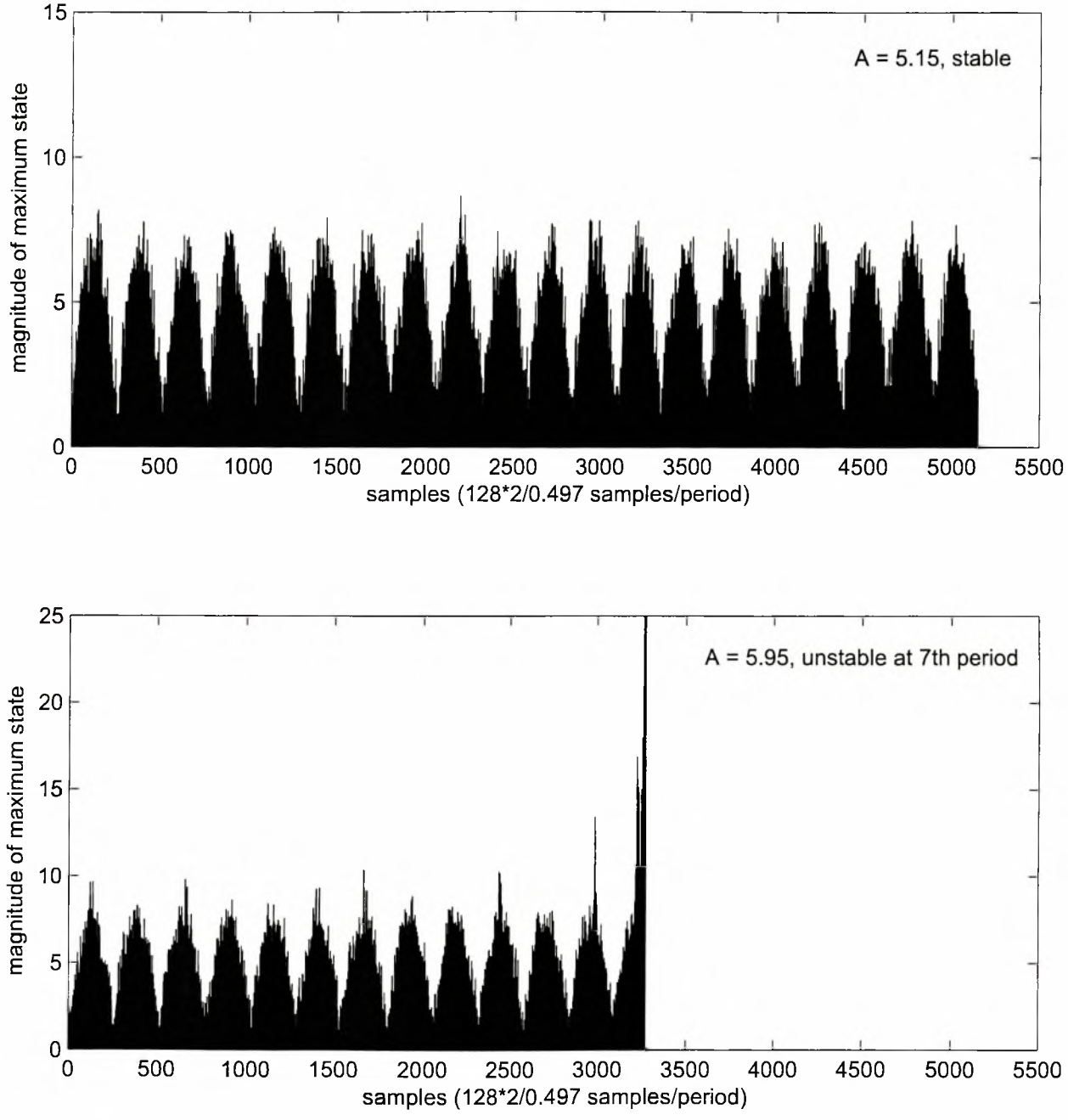

Figure 4.1: Maximum state for stable and unstable operation

Chapter 5

Conclusions

An oversampling modulator is a non-linear system, because of the quantization process taking place inside the loop. To some extend basic characteristics of low order modulators can be modeled and predicted with enough accuracy, using standard linear state space techniques. Such characteristics is the relationship between SNR and OSR as shown in the first chapter, or the relationship between the input amplitude and noise power for slow changing signals as studied in the second chapter. Parameters of decimation, as an intermediate state between the modulator output and low-pass filter can be roughly estimated, as seen in the third chapter.

In practice we want to push our design to maximum achievable specifications, by probably using a higher order modulator. However, it is almost impossible to predict theoretically its behavior in advance, using existing mathematical tools. This difficulty has been encountered in the fourth chapter when we tried to estimate with accuracy the stability margins of a third order modulator. There are regions that the modulator will remain stable for most cases, but for those that it does not, we can not specify the exact dynamics of the system that drive it to instability.

Since there is a great need for designing such converters, maybe this can lead to establishing a non-linear theory describing their dynamics. Through this work we show the weakness of existing mathematics to describe non-linear systems and also the power of these systems to provide very useful results.

Bibliography

[1] Bernhard E. Boser and Bruce A. Wooley, *The design of sigma delta modulation analog to digital converters*, IEEE Journal of Solid State Circuits **23** (1988), no. 6, 249–258.

[2] J. C. Candy, *A Use of Double Integration in Sigma Modulation*, IEEE Trans. Communications **33** (1985), no. 3, 249–258.

[3] James C. Candy, *Decimation for Sigma Delta Modulation*, IEEE Trans. Communications **34** (1986), no. 1, 72–76.

[4] James C. Candy and Oconnell J. Benjamin, *The Structure of Quantization Noise from Sigma-Delta Modulation*, IEEE Trans. Communications **29** (1981), no. 9, 1316–1326.

[5] C.-C. Hsu et al., *An 11b 800MS/s time-interleaved ADC with digital background calibration*, IEEE ISSCC Dig. Tech. Papers (2007), 464–465.

[6] K.Poulton et al., *A 20Gs/s 8b ADC with a 1MB memory in 0.18μm CMOS*, IEEE ISSCC Dig. Tech. Papers (2006), 264–265.

[7] S. Gupta, M. Choi, M. Inerfield, and J. Wang, *A 10Gs/s 11b time-interleaved ADC in 0.13μm CMOS*, IEEE ISSCC Dig. Tech. Papers (2003), 318–319.

[8] Ivar Lokken, Anders Vinje, Trond Sather, and Bjornar Hernes, *Quantizer Nonoverload Criteria in Sigma-Delta Modulators*, IEEE Trans. Communications **53** (2006), no. 12, 1383–1387.

[9] S. M. Louwsma, E. J. M. van Tuijl, M. Vertregt, and B. Nauta, *A 1.35GS/s, 10b, 175 mW time-interleaved AD converter in 0.13μm CMOS*, Symp. VLSI Circuits Dig. (2007), 62–63.

[10] V. Quiquempoix, P.Deval, A.Barreto, G.Bellini, J. Markus, J. Silva, and G. C. Temes, *A low-power 22-bit incremental ADC*, IEEE Journal of Solid State Circuits **41** (2006), no. 7.

[11] V. Quiquempoix, P.Deval, J. Markus, J. Silva, and G. C. Temes, *Incremental Delta-Sigma Structures for DC Measurement: an Overview*, IEEE 2006 Custom Intergrated Circuits Conference (CICC), 41–48.

[12] Gopal Raghavan, J. F. Jensen, J. Laskowski, M. Kardos, Michael G. Case, M. Sokolich, and Stephen Thomas, *Architecture, design, and test of continuous time tunable intermediate frequency bandpass delta sigma modulators*, IEEE Journal of Solid State Circuits **36** (2001), no. 1, 5–13.

[13] Richard Schreier and Gabor C.Temes, *Understanding Delta-Sigma Data Converters*, IEEE Press, New Jersey, 2005.

[14] Richard Schreier and Bo Zhang, *Delta sigma modulators employing continuous time circuitry*, IEEE Trans. Circuits and Systems **43** (1996), no. 4, 324–332.

[15] S.M.Louwsma, E.J.M.van Tuijl, M.Vertregt, and B.Nauta, *A 1.35GS/s, 10b, 175 mW time-interleaved AD converter in 0.13μm CMOS*, IEEE Journal of Solid State Circuits **43** (2008), no. 4, 778–786.

[16] www.dpreview.com, http://www.dpreview.com/reviews/pentaxk10d/.