

Auswirkungen der Quantisierung bei nichtlinearen Funktionen

W. Schlecker and H.-J. Pfeiderer

Abteilung Allgemeine Elektrotechnik und Mikroelektronik, Universität Ulm, Germany

Zusammenfassung. Um bei der Implementierung von digitalen Signalverarbeitungsalgorithmen den Hardwareaufwand zu minimieren, kann die Wortbreite der einzelnen Berechnungen optimiert werden. Dabei müssen die entstehenden Quantisierungsfehler berücksichtigt werden. In linear zeitinvarianten Systemen (LTI) können die Quantisierungsfehler in Abhängigkeit der statistischen Werte der Eingangssignale ermittelt werden. In vielen Algorithmen der digitalen Signalverarbeitung werden zusätzlich zu linearen Funktionen nichtlineare Funktionen verwendet. Dieser Beitrag befasst sich mit den Besonderheiten der Quantisierung bei Verwendung von nichtlinearen Funktionen. Dabei wird erläutert, welche Eigenschaften der Quantisierungsfehler am Eingang einer nichtlinearen Funktion haben sollte, um durch die Funktion die Verschlechterung des Signal-zu-Rauschleistungsverhältnisses (SNR) zu minimieren. Weiter wird der Fehler, der durch die Quantisierung nach einer nichtlinearen Funktion entsteht, am Beispiel der Multiplikation untersucht.

1 Einleitung

Hardwareimplementierungen auf FPGAs und ASICs erlauben es, die Wortbreite für die Berechnungen und Zwischenergebnisse frei zu wählen. Um die Verlustleistung und die benötigte Fläche zu minimieren und zugleich die Rechengenauigkeit zu maximieren, sind Modelle notwendig, die die Beziehung zwischen Wortbreite und Quantisierungsfehler darstellen. Viele der Untersuchungen zu Quantisierungseffekten befassen sich mit linear zeitinvarianten (LTI) Systemen (Taylor, 1983) und (Diniz et al., 2002). In Oppenheim and Schafer (1999) ist ein Quantisierungsmodell dargestellt, welches eine wertekontinuierliche Gleichverteilung für den Quantisierungsfehler annimmt. In

Correspondence to: W. Schlecker
(wolfgang.schlecker@uni-ulm.de)



Abbildung 1. Quantisierungsmodell

Constantinides et al. (1999) wurde das Modell mit einer wertediskreten Gleichverteilung für den Quantisierungsfehler verfeinert und es wurde gezeigt, dass dieses Modell für das Abschneiden weniger Stellen eines bereits quantisierten Signals genauer ist. In vielen Algorithmen der Nachrichtentechnik z.B. Teich et al. (2000) werden nichtlineare Funktionen wie die Multiplikation zweier Signale immer wichtiger. Constantinides (2003) und Constantinides et al. (2004) zeigen eine Methode, um die Eingangsfehler von nichtlinearen Funktionen anhand eines Kleinsignalmodells abzuschätzen. In diesem Bericht wollen wir auf die Besonderheiten bei nichtlinearen Funktionen und Unterschiede zu LTI-Systemen am Beispiel der Multiplikation eingehen. Weiter werden wir speziell die Auswirkung eines Eingangsmittelwerts auf den Ausgang bei LTI und nichtlinearen Systemen vergleichen.

2 Grundlagen zur Quantisierung

Im Weiteren werden die zwei gängigsten Methoden der Quantisierung von Zweierkomplementzahlen betrachtet: Abschneiden und Runden. Es soll auf n Nachkommastellen quantisiert werden, die Schrittweite des Quantisierers ist dann 2^{-n} . Das in Oppenheim and Schafer (1999) verwendete Quantisierungsmodell ersetzt die Quantisierung durch die Addition des Quantisierungsfehlers $e = y - x$ (Abb. 1). Dabei ist x das wertekontinuierliche Eingangssignal und y das Signal nach der Quantisierung. Für den Quantisierungsfehler werden dabei folgende statistische Annahmen getroffen:

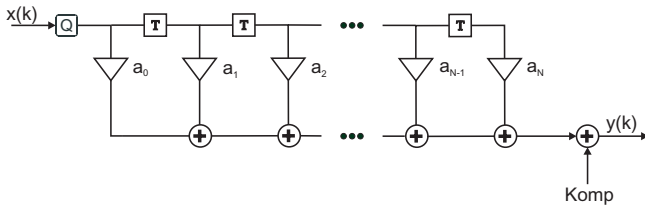


Abbildung 2. FIR Filter mit Quantisierung des Eingangssignals und Fehlermittelwertkompensation am Ausgang

- Der Fehler $e[k]$ ist die Abtastfolge eines stationären Zufallsprozesses.
- Der Fehler $e[k]$ ist unkorreliert mit dem Signal $x[k]$.
- Der Fehlerprozess lässt sich als weißes Rauschen beschreiben, d.h. die Zufallsvariablen des Fehlerprozesses sind unkorreliert.
- Die Wahrscheinlichkeitsdichtefunktion des Fehlers ist gleichverteilt über den Bereich des Quantisierungsfehlers.

Für “genügend komplexe” Signale, d.h. dass das Signal von Abtastwert zu Abtastwert mehrere Quantisierungsstufen überstreicht, sind diese Annahmen realistisch (Oppenheim and Schafer, 1999).

Ausgehend von dieser vereinfachten Modellannahme, der Gleichverteilung des Quantisierungsfehlers, kann nun die Fehlerfortpflanzung bei einfachen Rechenoperationen näher betrachtet werden. Der Mittelwert ist $\mu_e = -\frac{1}{2}2^{-n}$ für Abschneiden und $\mu_e = 0$ für Runden und die Varianz ist $\sigma_e^2 = \frac{1}{12}2^{-2n}$ für Abschneiden und für Runden.

Werden nur wenige Stellen einer bereits wertediskreten Zahl quantisiert, so ist dieses Fehlermodell ungenau. Für genauere statistische Werte des Fehlers werden alle Möglichkeiten (in diesem Fall endlich viele) gleichwahrscheinlich angenommen. Der Mittelwert berechnet sich aus der Summe aller möglichen abgeschnittenen Werte, dividiert durch die Anzahl der Möglichkeiten (Constantinides et al., 1999). Für eine Quantisierung eines Signals mit n_1 Nachkommastellen auf n_2 Nachkommastellen mit $n_1 > n_2$ wird der Mittelwert des Quantisierungsfehlers wie folgt berechnet. Für Abschneiden:

$$\mu_e = \frac{1}{2^{n_1-n_2}} \sum_{i=0}^{2^{n_1-n_2}-1} -i2^{-n_1} = -\frac{1}{2}(2^{-n_2} - 2^{-n_1})$$

für Runden:

$$\mu_e = \frac{1}{2^{n_1-n_2}} \sum_{i=0}^{2^{n_1-n_2}-1} (2^{-n_2-1} - i2^{-n_1}) = \frac{1}{2}2^{-n_1}.$$

Die Varianz wird analog dazu für Abschneiden und Runden bestimmt:

$$\sigma_e^2 = \frac{1}{12}(2^{-2n_2} - 2^{-2n_1}).$$

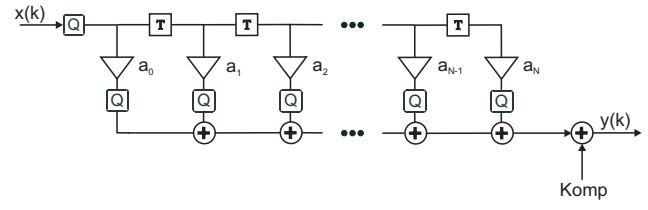


Abbildung 3. FIR Filter mit Quantisierung des Eingangssignals, Quantisierung der Skalierungsergebnisse und Kompensation des Fehlermittelwertes am Ausgang

Mit Hilfe dieser statistischen Werte kann das Signal-zu-Rauschleistungsverhältnis (SNR) des Quantisierungsrauschens berechnet werden. Das SNR ergibt sich aus dem Verhältnis von quadratischem Mittelwert des Signals zu dem quadratischen Mittelwert des Quantisierungsfehlers (Shanmugan and Breipohl, 1990):

$$SNR = \frac{\mu_{Sig}^2}{\mu_e^2} = \frac{\mu_{Sig}^2 + \sigma_{Sig}^2}{\mu_e^2 + \sigma_e^2} \quad (1)$$

3 Quantisierung in LTI-Systemen

In diesem Abschnitt betrachten wir die Quantisierung in LTI-Systemen am Beispiel eines Finite Impulse Response (FIR) Filters. Zunächst gehen wir von einem FIR-Filter aus, dessen Eingangssignal $x(k)$ quantisiert wird (Abb. 2). Das Ausgangssignal $y(k)$ berechnet sich folgendermaßen:

$$y(k) = \sum_{i=0}^N a_i x(k-i) \quad (2)$$

Damit lässt sich der Mittelwert des Quantisierungsfehlers e_y am Ausgang folgendermaßen berechnen:

$$\mu_{ey} = \mu_{ex} \sum_{i=0}^N a_i. \quad (3)$$

Wird das wertekontinuierliche Eingangssignal $x(k)$ auf n Nachkommastellen abgeschnitten, ist der Fehlermittelwert am Ausgang

$$\mu_{ey} = -\frac{1}{2}2^{-n} \sum_{i=0}^N a_i \quad (4)$$

und für Runden

$$\mu_{ey} = 0. \quad (5)$$

Die Varianz des Fehlers am Ausgang ist für Runden und Abschneiden

$$\sigma_{ey}^2 = \sigma_{ex}^2 \sum_{i=0}^N a_i^2 = \frac{1}{12}2^{-2n} \sum_{i=0}^N a_i^2 \quad (6)$$

Dies bedeutet, dass sich Runden und Abschneiden im Fehler nur durch den Fehlermittelwert unterscheiden. Bei Abschneiden am Eingang und Subtraktion des vorhersagbaren Fehlermittelwerts (4) am Ausgang wird derselbe Fehlermittelwert wie bei Runden am Eingang erreicht. In diesem Fall bringt dies aber keine Hardwareeinsparung, da Runden (eine Addition) am Eingang durch eine Addition am Ausgang ersetzt wird. Betrachten wir nun ein FIR-Filter, bei dem nach den einzelnen Skalierungen ebenfalls quantisiert wird (Abb. 3). Mit der Annahme, dass die quantisierten Stellen gleichverteilt sind und viele Stellen quantisiert werden, so dass der Fehler mit einer kontinuierlichen Gleichverteilung genähert werden kann, ist der Mittelwert des Ausgangssignals für Abschneiden:

$$\mu_{ey} = -\frac{1}{2}2^{-n} \left(N + 1 + \sum_{i=0}^N a_i \right). \quad (7)$$

Wird gerundet, so ist $\mu_{ey} = 0$. Die Varianz des Fehlers am Ausgang ist für Runden und Abschneiden gleich:

$$\sigma_{ey}^2 = \frac{1}{12}2^{-2n} \left(N + 1 + \sum_{i=0}^N a_i^2 \right). \quad (8)$$

Auch hier kann der Mittelwert kompensiert werden, so dass mit Abschneiden das gleiche SNR wie bei Runden erreicht wird. Haben die Skalierungsfaktoren a_i nur wenige Nachkommastellen $n_{a,i}$, so gilt die Annahme, dass der Fehler kontinuierlich ist, nicht mehr. Der Fehlermittelwert am Ausgang berechnet sich für Abschneiden dann folgendermaßen:

$$\begin{aligned} \mu_{ey} &= -\frac{1}{2}2^{-n} \sum_{i=0}^N a_i - \frac{1}{2} \sum_{i=0}^N (2^{-n} - 2^{-n-n_{a,i}}) \\ &= -\frac{1}{2}2^{-n} \sum_{i=0}^N (a_i + 1 - 2^{-n_{a,i}}). \end{aligned}$$

Für Runden ist der Fehlermittelwert dann

$$\mu_{ey} = \frac{1}{2} \sum_{i=0}^N (2^{-n-n_{a,i}}).$$

Die Varianz ist für Runden und Abschneiden:

$$\begin{aligned} \sigma_{ey}^2 &= \frac{1}{12}2^{-2n} \sum_{i=0}^N a_i^2 + \frac{1}{12} \sum_{i=0}^N (2^{-2n} - 2^{-2n-2n_{a,i}}) \\ &= \frac{1}{12}2^{-2n} \sum_{i=0}^N (a_i^2 + 1 - 2^{-2n_{a,i}}). \end{aligned}$$

Abbildung 4 zeigt das simulierte SNR aufgetragen über die Anzahl der Nachkommastellen für Runden, für Abschneiden und für Abschneiden mit Mittelwertkompensation. Für die Simulation wurde von Eingangswerten $x(k)$ ausgegangen, die im Intervall $[-1; 1]$ gleichverteilt sind. Wie man erkennt, erreicht man für Abschneiden mit Mittelwertkompensation das gleiche SNR wie für Runden. Allgemein

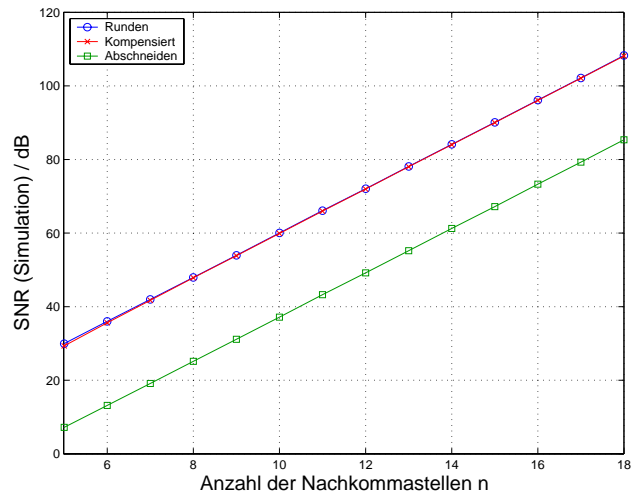


Abbildung 4. SNR des FIR-Filters (Abb. 3) mit Quantisierung am Eingang für Runden, Abschneiden und Abschneiden mit Kompensation des Ausgangssignals

kann gesagt werden, dass Runden in LTI-Systemen nicht notwendig ist, solange die quantisierten Stellen gleichverteilt sind, da die Varianz des Fehlers σ_e^2 in beiden Fällen gleich ist. Der Unterschied macht sich nur in dem Fehlermittelwert μ_e bemerkbar und kann durch nur eine Addition bzw. Subtraktion eliminiert werden.

Dabei stellt sich die Frage, wann die Bedingung, dass die quantisierten Stellen diskret gleichverteilt sind, erfüllt ist. Sind die letzten m Stellen des quantisierten Eingangssignals gleichverteilt, so sind auch die letzten m Stellen aller Skalierungsergebnisse und Additionsergebnisse gleichverteilt. Dabei dürfen nur Stellen, die ungleich Null werden können, berücksichtigt werden. Dies kann folgendermaßen gezeigt werden:

u sei eine 1 Bit-Zahl, die diskret gleichverteilt ist, mit den Wahrscheinlichkeiten $P(u=1) = P(u=0) = 0,5$, v sei ebenfalls eine 1 Bit-Zahl mit $P(v=1) = \gamma$, $P(v=0) = 1 - \gamma$. Werden die Zahlen u, v zu einer 1 Bit-Zahl w addiert ($w = u + v$), so ist die Wahrscheinlichkeit:

$$\begin{aligned} P(w = 1) &= P(u = 1) \cdot P(v = 0) + P(u = 0) \cdot P(v = 1) \\ &= P(u = 1) \cdot \gamma + P(u = 0) \cdot (1 - \gamma) \\ &= 0,5 \end{aligned}$$

Da die Addition von mehreren Stellen auf 1 Bit-Additionen zurückgeführt werden kann, gilt für das Additionsergebnis zweier Zahlen, dass die letzten m Stellen unabhängig voneinander und mit Wahrscheinlichkeit 0,5 Eins sind, also gleichverteilt, wenn die letzten m Stellen eines Summanden unabhängig voneinander und mit der Wahrscheinlichkeit 0,5 Eins sind.

Die Skalierung wiederum kann auf einzelne Additionen zurückgeführt werden. Sind die letzten m Bit von x gleich-

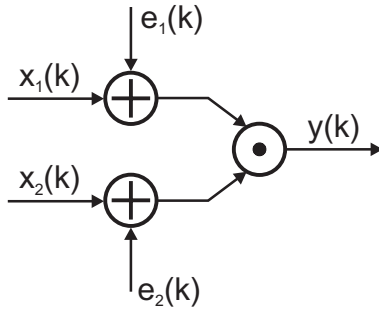


Abbildung 5. Multiplikation mit Quantisierung der Eingangssignale

verteilt, so sind die letzten m Bit des Summanden, der durch die Multiplikation von x mit $a_{i,0} = 1$ entsteht, gleichverteilt und damit auch die letzten m Bit des Skalierungsergebnisses.

4 Quantisierung in nichtlinearen Systemen am Beispiel der Multiplikation

Nun betrachten wir ein nichtlineares System am Beispiel der Multiplikation. Im Gegensatz zu den Skalierungen im vorherigen Abschnitt bilden wir das Produkt von zwei unabhängigen Zufallszahlen. Zunächst betrachten wir die Auswirkung der Quantisierung der Eingangssignale auf n_x Nachkommastellen auf das Ausgangssignal $y(k)$ der Multiplikation (Abb. 5).

$$y = (x_1 + e_1) \cdot (x_2 + e_2) \\ = x_1 \cdot x_2 + x_1 \cdot e_2 + x_2 \cdot e_1 + e_1 \cdot e_2$$

mit dem Quantisierungsfehler

$$e_{mult} = x_1 \cdot e_2 + x_2 \cdot e_1 + e_1 \cdot e_2$$

Sind die Eingangssignale statistisch unabhängig, so ist der Mittelwert des Quantisierungsfehlers:

$$\mu_{e,mult} = \mu_{x1} \cdot \mu_{e2} + \mu_{x2} \cdot \mu_{e1} + \mu_{e1} \cdot \mu_{e2}$$

Die Varianz beträgt:

$$\sigma_{e,mult}^2 = \mu_{x12} \mu_{e22} - (\mu_{x1} \mu_{e2})^2 + \mu_{x22} \mu_{e12} - (\mu_{x2} \mu_{e1})^2 \\ + \mu_{e12} \mu_{e22} - (\mu_{e1} \mu_{e2})^2$$

Mit der Annahme, dass die Eingangssignale mittelwertfrei sind ($\mu_{x1} = \mu_{x2} = 0$), berechnen sich Mittelwert und Varianz zu

$$\mu_{e,mult} = \mu_{e1} \cdot \mu_{e2} \\ \sigma_{e,mult}^2 = \sigma_{x1}^2 \sigma_{e2}^2 + \sigma_{x2}^2 \sigma_{e1}^2 + \sigma_{e1}^2 \sigma_{e2}^2 \\ + \sigma_{x1}^2 \mu_{e2}^2 + \sigma_{x2}^2 \mu_{e1}^2 + \sigma_{e1}^2 \mu_{e2}^2 + \sigma_{e2}^2 \mu_{e1}^2.$$

Das bedeutet, dass die Fehlermittelwerte (μ_{e1}, μ_{e2}) die Varianz des Fehlers am Ausgang erhöhen. Daraus folgt, dass

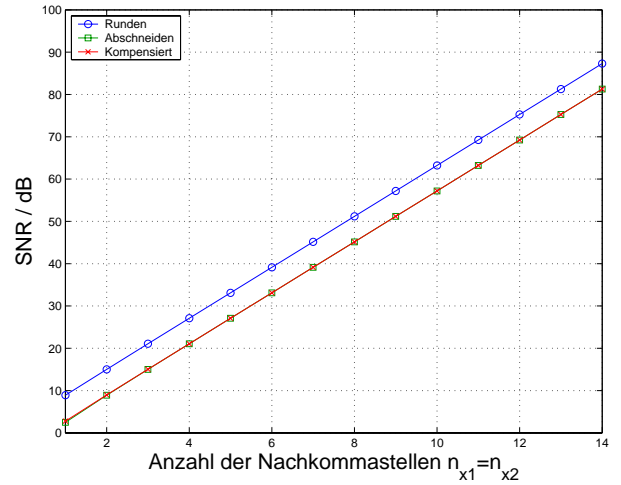


Abbildung 6. SNR am Ausgang der Multiplikation für Runden, Abschneiden und Abschneiden mit Kompensation des Fehlermittelwertes am Ausgang

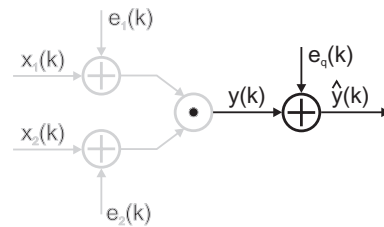


Abbildung 7. Quantisierung des Ausgangssignals der Multiplikation mit $n_{x1} + n_{x2}$ Nachkommastellen auf n_y Nachkommastellen

der Fehlermittelwert der z.B. beim Abschneiden entsteht nicht wie bei LTI-Systemen am Ausgang kompensiert werden kann. Abbildung 6 zeigt das simulierte SNR des Ausgangssignals der Multiplikation aufgetragen über die Anzahl der Nachkommastellen der Eingangssignale für Runden bzw. Abschneiden am Eingang und für Abschneiden am Eingang mit Kompensation des Mittelwertes am Ausgang. Für die Eingangssignale wurde eine Gleichverteilung im Intervall $[-1; 1]$ angenommen. Man erkennt, dass das SNR für Abschneiden trotz der Kompensation des Fehlermittelwertes am Ausgang 6 dB schlechter ist. Allgemein kann gesagt werden, dass ein Fehlermittelwert am Eingang einer nichtlinearen Funktion zu einer Vergrößerung der Fehlervarianz am Ausgang führen kann. Deshalb sollte der Fehlermittelwert der Eingangssignale Null sein. Liegt ein wertekontinuierliches Signal vor, kann dies durch Runden erreicht werden. Wird das Eingangssignal mit einem LTI-System berechnet, so sollte der Fehlermittelwert wie im vorherigen Abschnitt beschrieben kompensiert werden.

Betrachten wir nun die Quantisierung des Ausgangssignals der Multiplikation $y(k)$ mit $n_{x1} + n_{x2}$ Nachkommastellen

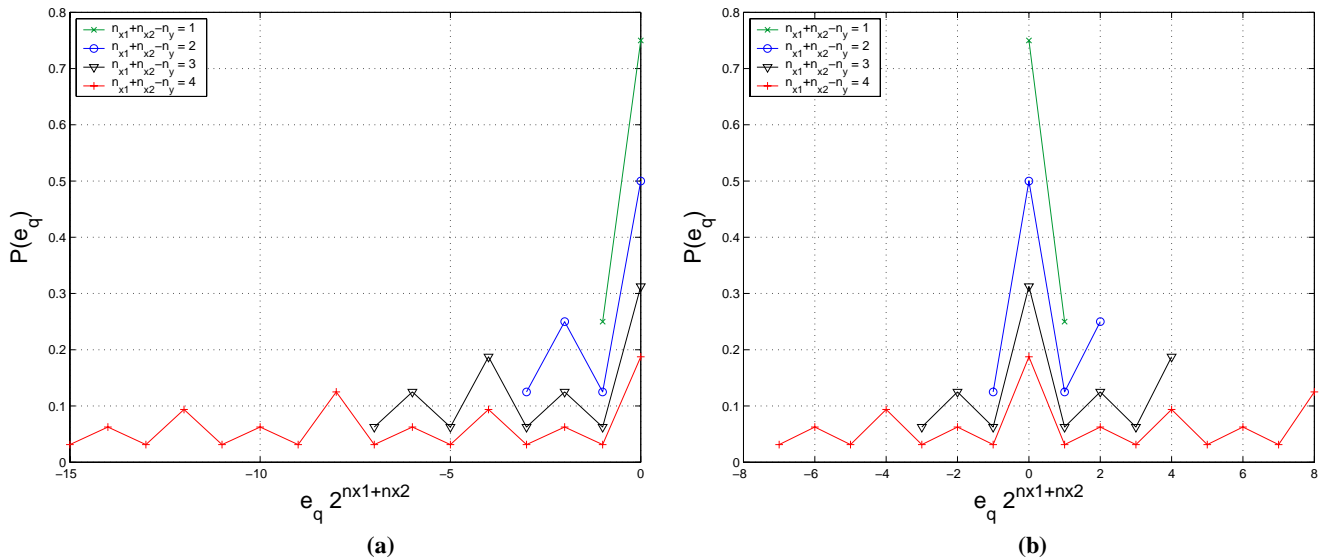


Abbildung 8. Wahrscheinlichkeitsfunktion des Quantisierungsfehlers e_q (a) beim Abschneiden (b) beim Runden

stellen auf n_y Nachkommastellen (Abb. 7). Dieser Fehler ist im Gegensatz zu dem Quantisierungsfehler in LTI-Systemen nicht mehr gleichverteilt. Bei gleichverteilten Eingangssignalen ist z.B. die Wahrscheinlichkeit P , dass das Least Significant Bit (LSB) des Ausgangssignals Eins ist:

$$P(\text{LSB}_{\text{OUT}} = 1) = P(\text{LSB}_{\text{IN}} = 1) \cdot P(\text{LSB}_{\text{IN}} = 1) = 0,5 \cdot 0,5 = 0,25$$

In Abb. 8 sind die Wahrscheinlichkeiten für die Quantisierungsfehler e_q aufgetragen unter der Annahme, dass die Anzahl der quantisierten Stellen des Ausgangssignals bei den Eingangssignalen diskret gleichverteilt sind. Z.B. ist $n_{x1} + n_{x2} - n_y = 3$, so wird angenommen, dass die drei letzten Bits beider Eingangssignale diskret gleichverteilt sind. Diese von einer diskreten Gleichverteilung stark abweichenden Wahrscheinlichkeiten führen dazu, dass der Mittelwert und der quadratische Mittelwert des Fehlers e_q deutlich von Mittelwert und quadratischem Mittelwert eines gleichverteilten Fehlers abweichen.

Für Runden gilt:

$$\mu_{eq} = \frac{1}{4}(n_{x1} + n_{x2} - n_y)2^{-n_{x1}-n_{x2}}$$

$$\mu_{eq^2} = \frac{1}{12}2^{-2n_y} - \frac{1}{12}2^{-2n_{x1}-2n_{x2}}$$

Für Abschneiden gilt:

$$\mu_{eq} = \frac{1}{2}2^{-n_{x1}-n_{x2}}$$

$$+ \frac{1}{4}(n_{x1} + n_{x2} - n_y)2^{-n_{x1}-n_{x2}} - \frac{1}{2}2^{-n_y} \quad (9)$$

$$\mu_{eq^2} = -\frac{1}{4}2^{-n_{x1}-n_{x2}-n_y}$$

$$- \frac{1}{4}(n_{x1} + n_{x2} - n_y)2^{-n_{x1}-n_{x2}-n_y}$$

$$+ \frac{1}{3}2^{-2n_y} - \frac{1}{12}2^{-2n_{x1}-2n_{x2}}$$

In Abb. 9 sind die theoretischen Mittelwerte und quadratischen Mittelwerte für e_q und diskreter Gleichverteilung aufgetragen. Werden nur wenige Stellen quantisiert, so weichen die Mittelwerte von e_q und diskreter Gleichverteilung stark voneinander ab. Werden viele Stellen abgeschnitten, so sind Abweichungen vernachlässigbar.

5 Zusammenfassung

In dem vorliegenden Beitrag wurden die Auswirkungen von Quantisierung in LTI-Systemen und vor und nach der Multiplikation von zwei Zufallszahlen untersucht. Dabei wurde festgestellt, dass in LTI-Systemen Abschneiden im Vergleich zu Runden lediglich zu einer Vergrößerung des Fehlermittelwertes am Ausgang führt, der vorhersagbar ist. Dies erlaubt es, den Fehlermittelwert zu kompensieren und damit das gleiche SNR wie für Runden zu erreichen. Bei der Multiplikation einer nichtlinearen Funktion haben wir gesehen, dass der Fehlermittelwert am Eingang zu einer Vergrößerung der Fehlervarianz und zu einer Verschlechterung des SNR am Ausgang führt. Diese Verschlechterung kann durch Kompensation des Mittelwertes am Ausgang nicht mehr rückgängig gemacht werden. Allgemein kann gesagt werden, dass ein Fehlermittelwert am Eingang von nichtlinearen Funktionen zu einer Verschlechterung des SNR führen kann und deshalb der Fehlermittelwert am Eingang kompensiert werden sollte.

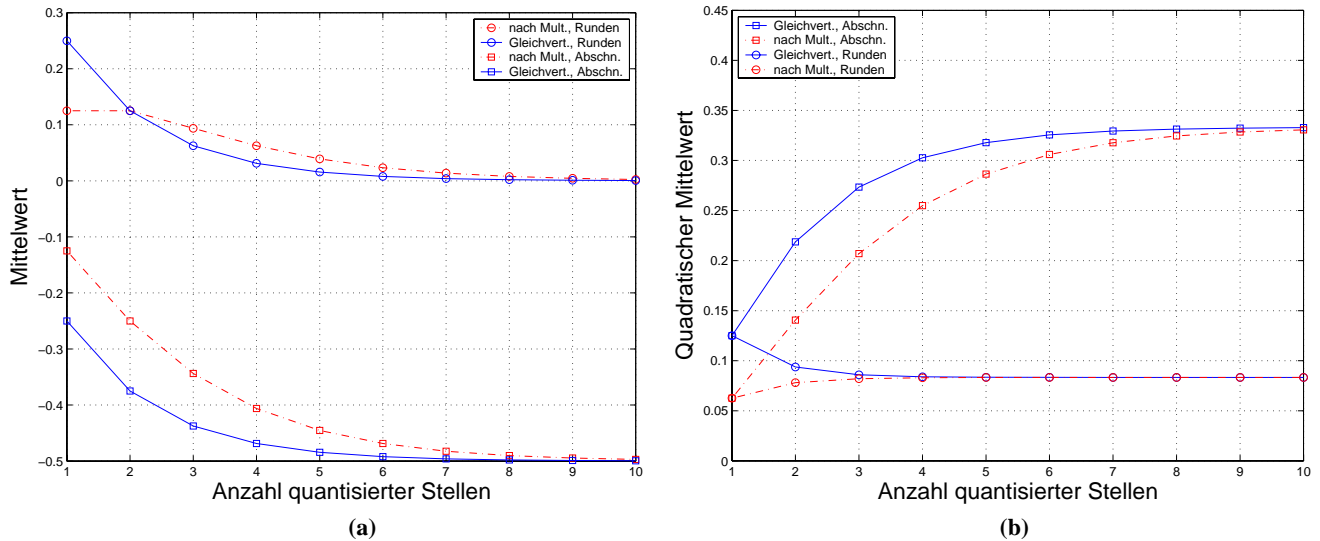


Abbildung 9. Vergleich des (a) Mittelwert und (b) quadratischer Mittelwert von e_q und gleichverteiltem Fehler

Weiter wurde gezeigt, dass der Quantisierungsfehler am Ausgang der Multiplikation trotz der gleichverteilten Eingangssignale stark von der diskreten Gleichverteilung abweicht. Dies führt zu Mittelwerten und einer Varianz, die bei der Quantisierung von wenigen Stellen stark von den Mittelwerten der diskreten Gleichverteilung abweichen.

Literatur

- Constantinides, G. A.: Perturbation Analysis for Word-length Optimization, Proceedings of the 11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, 2003.
- Constantinides, G. A., Cheung, P. Y. K., and Luk, W.: Truncation noise in fixed-point SFGs, IEE Electronics Letters, 35, 2012–2014, 1999.
- Constantinides, G. A., Cheung, P. Y. K., and Luk, W.: Synthesis and Optimization of SDSP Algorithms, KLUWER ACADEMIC PUBLISHERS, Boston, Dordrecht, London, 2004.
- Diniz, P., Netto, S., and Silva, E. D.: Digital Signal Processing: System Analysis and Design, Cambridge University Press, New York, NY, USA, 2002.
- Oppenheim, A. V. and Schaffer, R. W.: Zeitdiskrete Signalverarbeitung, vol. 3, R. Oldenburg Verlag, Rosenheimerstraße 145, D-71671 München, 1999.
- Shanmugan, K. S. and Breipohl, A. B.: Random Signals; Detection, Estimation and Data Analysis, John Wiley and Sons, New York, 1990.
- Taylor, F. J.: Digital Filter Design Handbook, Marcel Dekker, Inc., New York, NY, USA, 1983.
- Teich, W. T., Engelhart, A., Schlecker, W., Gessler, R., and Pfeiderer, H.-J.: Towards an Efficient Hardware Implementation of Recurrent Neural Network Based Multiuser Detection, IEEE 6th Int. Symp. on Spread-Spectrum Tech. and Appl., 2000.