Advances in Radio Science, 3, 343–347, 2005
SRef-ID: 1684-9973/ars/2005-3-343

Advances in
Radio Science

Analytische Betrachtung des Quantisierungsfehlers bei grundlegenden Rechenoperationen der digitalen Signalverarbeitung

W. Schlecker, C. Beuschel, and H.-J. Pfleiderer

Universität Ulm, Abt. Allgemeine Elektrotechnik und Mikroelektronik, 89081 Ulm, Albert-Einstein-Allee 43, Germany

Zusammenfassung. Bei der Realisierung digitaler Schaltungen mit einem ASIC oder FPGA kann die Wortbreite der Berechnungen frei gewählt werden. Um die Fläche bzw. die benötigte Hardware zu minimieren, wird versucht, die Berechnungen mit minimaler Wortbreite zu implementieren. Dabei muss jedoch das Quantisierungsrauschen, das durch das Beschränken der Wortbreite entsteht, berücksichtigt werden. Häufig wird hierzu eine worst-case Abschätzung des Fehlers gemacht oder mit Simulationen die benötigte Wortbreite bestimmt. Der vorliegende Beitrag betrachtet die Auswirkungen der Quantisierung analytisch. Dabei wird von gleichverteilten Eingangssignalen ausgegangen. Es wird das Quantisierungsrauschen in Abhängigkeit von der Eingangs- und Ausgangswortbreite bei der Multiplikation und beim Skalarprodukt betrachtet. Die Untersuchungen wurden für Runden und für Abschneiden analytisch durchgeführt und durch Simulation bestätigt.

1 Einleitung

Immer mehr digitale Signalverarbeitungsalgorithmen werden auf Field Programmable Gate Arrays (FPGAs), Application Specific Integrated Circuits (ASICs) und Digital Signal Processors (DSPs) implementiert, dabei wächst die Komplexität der Algorithmen in gleichem Maße wie die Komplexität der Bausteine. FPGAs und ASICs haben gegenüber DSPs den Vorteil, dass die Genauigkeit der internen Berechnung beliebig gewählt werden kann. Um Hardware (HW) Kapazität, Configurable Logic Blocks (CLBs) beim FPGA oder Fläche beim ASIC zu sparen, sollten die Wortbreiten der internen Berechnungen möglichst klein gewählt werden. Ein weiterer Vorteil kleinerer Wortbreiten sind auch eine höhere maximale Taktfrequenz und ein geringerer Leistungsverbrauch.

Zunächst gibt es zwei grundsätzliche Möglichkeiten, die Wortbreiten der einzelnen Berechnungen zu wählen: Alle internen Wortbreiten sind gleich, oder die Wortbreiten der einzelnen Berechnungen sind unterschiedlich. Bei einem DSP werden z.B. alle internen Berechnungen mit der gleichen Wortbreite durchgeführt. FPGAs und ASICs dagegen erlauben es, die einzelnen Berechnungen mit unterschiedlichen Wortbreiten durchzuführen. Dabei ist zu beachten, dass durch geringere Rechengenauigkeit Quantisierungsfehler entstehen. Nun ist die Frage, wie man die Wortbreiten ermitteln kann, bei denen der Quantisierungsfehler tolerabel ist. Es gibt die Möglichkeit, die Schaltung so auszulegen, dass keine zusätzlichen Quantisierungsfehler entstehen. Dies eignet sich allerdings nur für kleine Schaltungen, da die Wortbreiten von Berechnung zu Berechnung immer größer werden. Eine weitere Möglichkeit ist die Simulation der Gesamtschaltung mit verschiedenen internen Wortbreiten, um so die minimalen Wortbreiten zu ermitteln, für die der Quantisierungsfehler tolerabel ist. Das Problem dabei ist die Simulationszeit bei komplexen Schaltungen und die vielen Quantisierungsmöglichkeiten. Weiter stellt sich häufig die Frage, ob Runden oder Abschneiden zum Quantisieren verwendet werden soll. In diesem Beitrag wird der Quantisierungsfehler analytisch betrachtet, um Aussagen über die Auswirkungen der Quantisierung ableiten zu können und den Unterschied zwischen Runden und Abschneiden genauer zu untersuchen.

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

2 Grundlagen zur Quantisierung

Im folgenden werden die zwei gängigsten Methoden der Quantisierung (Abschneiden und Runden) betrachtet. Die Schrittweite des Quantisierers sei Δ. Beim Abschneiden wird dem Eingangswert das nächstkleinere ganzzahlige Vielfache von Δ zugewiesen. Ist der Eingangswert bereits ein ganzzahliges Vielfaches von Δ, so wird dieser Wert zugewiesen. Für Zweierkomplementzahlen (K2-Zahlen) bedeutet dies nichts anderes, als dass alle Stellen nach der letzten Stelle, die behalten werden soll, weggelassen werden.

Beim Runden wird dem Eingangswert das nächstliegende ganzzahlige Vielfache von Δ zugewiesen. Ist der Abstand zum nächsten kleineren und größeren Wert gleich, so wird die Zahl wie beim kaufmännischen Runden dem größeren Wert zugewiesen. Dies bedeutet für Runden bei K2-Zahlen, dass zuerst an der Stelle hinter der letzten Stelle eine Eins addiert wird und danach abgeschnitten wird.

Abbildung 1. Quantisierungsmodell.

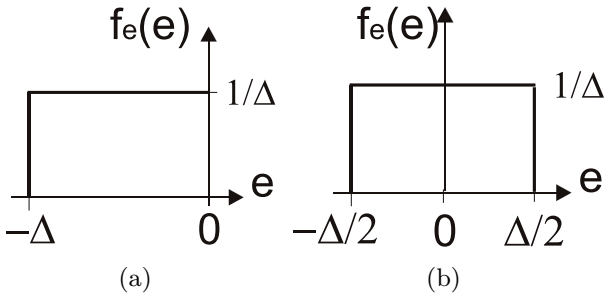

Abbildung 2. Verteilungsdichtefunktion des Quantisierungsfehlers **(a)** beim Abschneiden **(b)** beim Runden.

In der Literatur (Oppenheim and Schafer, 1999) ist folgendes statistisches Modell für die Quantisierung zu finden (siehe Abbildung 1): Die Quantisierung wird ersetzt durch die Addition des Quantisierungsfehlers $e{=}\hat{x}{-}x$. Dabei ist x das kontinuierliche Eingangssignal und $\hat{x}$ das quantisierte Ausgangssignal.

Für den Quantisierungsfehler gelten dabei folgende statistische Annahmen:

- Der Fehler $e[n]$ ist die Abtastfolge eines stationären Zufallsprozesses.
- Der Fehler $e[n]$ ist unkorreliert mit dem Signal $x[n]$.
- 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 (Abbildung 2a und b).

Natürlich kann man nicht bei jedem Signal davon ausgehen, dass diese vereinfachenden Annahmen für den Quantisierungsfehler auch gelten. Es kann allerdings gezeigt werden, dass sie zutreffen, sobald das Signal „genügend komplex“ ist, d.h. sobald das Signal von Abtastwert zu Abtastwert mehrere Quantisierungsstufen überstreicht (Oppenheim and Schafer, 1999). Mit dieser vereinfachten Modellannahme für den Quantisierungsfehler kann nun die Fehlerfortpflanzung bei einfachen Rechenoperationen näher betrachtet werden. Ausgehend von der Gleichverteilung des Quantisierungsfehlers ist der Mittelwert $\mu_e{=}{-}\frac{\Delta}{2}$ für Abschneiden und $\mu_e{=}0$ für Runden und die Varianz $\sigma_e{=}\frac{1}{12}\Delta^2$ für Abschneiden und für Runden.

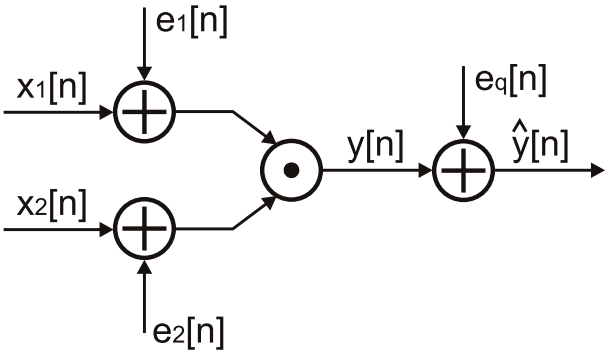

Abbildung 3. Quantisierungsmodell der Mutliplikation.

Dieses Modell für den Quantisierungsfehler trifft nur zu, wenn viele Stellen abgeschnitten oder gerundet werden. Handelt es sich um nur wenige Stellen, 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 von B1 Nachkommastellen auf B2 Nachkommmastellen wird der Mittelwert des Quantisierungsfehlers wie folgt berechnet. Für Abschneiden:

$$\mu_e = \frac{1}{2^{\mathrm{B1-B2}}} \sum_{i=0}^{2^{\mathrm{B1-B2}}} i2^{-\mathrm{B1}} = -\frac{1}{2}(2^{-\mathrm{B2}} - 2^{-\mathrm{B1}}) \tag{1}$$

für Runden:

$$\mu_e = \frac{1}{2^{\mathrm{B1-B2}}} \sum_{i=0}^{2^{\mathrm{B1-B2}}} (i2^{-\mathrm{B1}} - 2^{-\mathrm{B1}}) = -\frac{1}{2}2^{-\mathrm{B2}} \tag{2}$$

Die Varianz wird analog dazu für Abschneiden und Runden bestimmt:

$$\sigma_e^2 = \frac{1}{12}(2^{-2\mathrm{B2}} - 2^{-2\mathrm{B1}}) \tag{3}$$

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

$$\mathrm{SNR} = \frac{\mu_{Sig^2}}{\mu_{e^2}} = \frac{\mu_{Sig}^2 + \sigma_{Sig}^2}{\mu_e^2 + \sigma_e^2} \tag{4}$$

3 Quantisierung bei der Multiplikation

Mit dem Quantisierungmodell aus dem vorhergehenden Abschnitt werden nun die Auswirkungen der Quantisierung bei der Multiplikation untersucht. Im Gegensatz zu den Multiplikationen bzw. Skalierungen in linearen Systemen

wollen wir das Produkt von zwei unabhängigen Zufallszahlen, eine nichtlineare Funktion, betrachten. Wir gehen von zwei Zufallszahlen aus, die quantisiert werden, und anschließend multipliziert werden. Das Produkt dieser Zahlen wird wiederum quantisiert (siehe Abbildung 3). Betrachten wir zunächst nur die Quantisierungen vor der Multiplikation und die Auswirkungen der Quantisierungsfehler auf das Produkt. x_1 und x_2 sind unsere Zufallsgrößen, e_1 und e_2 sei der jeweilige Quantisierungsfehler, $\hat{y}$ das Produkt der quantisierten Zufallszahlen:

$$\hat{y} = (x_1+e_1)\cdot(x_2+e_2) = x_1\cdot x_2+x_1\cdot e_2+x_2\cdot e_1+e_1\cdot e_2 \quad (5)$$

Daraus ergibt sich ein Quantisierungsfehler von:

$$\hat{y} - y = x_1\cdot e_2 + x_2\cdot e_1 + e_1\cdot e_2 \quad (6)$$

Den Mittelwert bzw. quadratischen Mittelwert eines Produkts erhält man, indem die Mittelwerte bzw. die quadratischen Mittelwerte der Faktoren multipliziert werden (Beichelt, 1995). Damit ergibt sich ein Mittelwert des Fehlers von:

$$\mu_{e,mult} = \mu_{x1}\cdot\mu_{e2} + \mu_{x2}\cdot\mu_{e1} + \mu_{e1}\cdot\mu_{e2} \quad (7)$$

und ein quadratischer Mittelwert von:

$$\mu_{e,mult^2} = \mu_{x1^2}\cdot\mu_{e2^2} + \mu_{x2^2}\cdot\mu_{e1^2} + \mu_{e1^2}\cdot\mu_{e2^2} + 2\cdot(\mu_{x1}\cdot\mu_{x2}\cdot\mu_{e1}\cdot\mu_{e2} + \mu_{x1}\cdot\mu_{e1}\cdot\mu_{e2}^2 + \mu_{x2}\cdot\mu_{e1}^2\cdot\mu_{e2}) \quad (8)$$

Wird die Quantisierung nach der Multiplikation mit berücksichtigt, so erhält man folgenden Mittelwert und quadratischen Mittelwert für den Gesamtfehler:

$$\mu_{e,ges} = \mu_{x1}\cdot\mu_{e2} + \mu_{x2}\cdot\mu_{e1} + \mu_{e1}\cdot\mu_{e2} + \mu_{eq} \quad (9)$$

$$\mu_{e,mult^2} = \mu_{x1^2}\cdot\mu_{e2^2} + \mu_{x2^2}\cdot\mu_{e1^2} + \mu_{e1^2}\cdot\mu_{e2^2} + \mu_{eq^2} + 2\cdot\Big(\mu_{x1}\cdot\mu_{x2}\cdot\mu_{e1}\cdot\mu_{e2} + \mu_{x1}\cdot\mu_{e1}\cdot\mu_{e2}^2 + \mu_{x2}\cdot\mu_{e1}^2\cdot\mu_{e2} + \mu_{x1}\cdot\mu_{e2}\cdot\mu_{eq} + \mu_{x2}\cdot\mu_{e1}\cdot\mu_{eq} + \mu_{e1}\cdot\mu_{e2}\cdot\mu_{eq}\Big), \quad (10)$$

wobei e_q die Quantisierung nach der Multiplikation beschreibt. Im folgenden wählen wir für die Eingangssignale mittelwertfreie, gleichverteilte Zufallszahlen. Damit ergibt sich ein SNR von:

$$\mathrm{SNR} = \frac{\mu_{x1^2}\cdot\mu_{x1^2}}{\mu_{x1^2}\cdot\mu_{e2^2} + \mu_{x2^2}\cdot\mu_{e1^2} + \mu_{e1^2}\cdot\mu_{e2^2} + \mu_{eq^2}} \quad (11)$$

In Abbildung 4 ist das berechnete SNR durch den Quantisierungsfehler aufgetragen über der Wortbreite der Quantisierung nach der Multiplikation. Die Eingangswerte sind gleichverteilt im Intervall $]-1;1[$ und werden vor der Multiplikation auf BE Nachkommastellen gerundet. Simulationen derselben Konstellationen ergeben dieselbe Kurve, sofern die Eingangswortbreiten nicht zu klein gewählt werden (>5 Bit). In diesem Diagramm erkennt man, dass wenn die Ausgangswortbreite kleiner ist als die Eingangswortbreite, das SNR für Runden 6 dB besser ist als für Abschneiden; weiter gilt in diesem Bereich, dass für jedes weitere Bit das SNR um 6 dB besser wird. Das heißt, man kann eine Quantisierung durch Runden auf n Bit ersetzen durch aufwandsgünstigeres Abschneiden auf $n+1$ Bit und erhält dasselbe SNR.

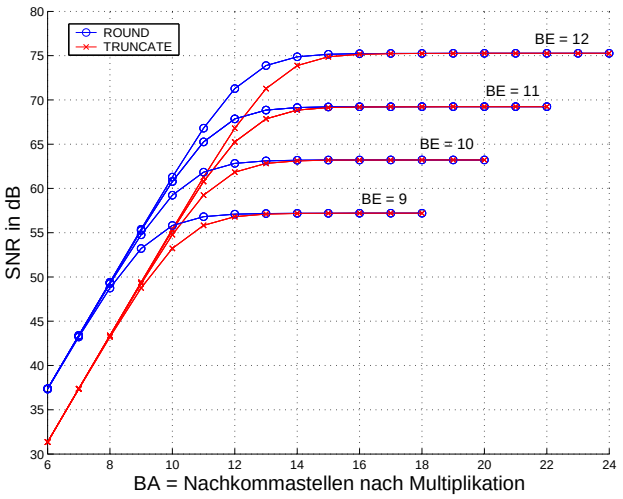

Abbildung 4. SNR über Wortbreite am Ausgang für Abschneiden und für Runden (Eingang jeweils auf BE Nachkommastellen gerundet).

Weiter kann man in Abbildung 4 erkennen, dass das SNR in eine Sättigung geht. Eine Erhöhung der Wortbreite am Quantisierer nach der Multiplikation auf mehr als BE+2 Nachkommastellen für Runden bzw. BE+3 Bit für Abschneiden bringt keine signifikante Verbesserung des SNR. Dieser Umstand lässt sich bei anderen Eingangswortbreiten ebenfalls beobachten. Ist die Wortbreite nach der Multiplikation bei Runden 2, bei Abschneiden 3 Bit größer als die Eingangswortbreite, so erhält man durch weitere Erhöhung der Wortbreite nur einen minimalen SNR-Gewinn.

4 Quantisierung beim Skalarprodukt

Nachdem die Auswirkung der Quantisierung bei der Multiplikation betrachtet wurde, wollen wir nun die Auswirkungen bei dem Skalarprodukt von 2 Zufallsvektoren untersuchen. Dabei wurde von einer Quantisierung der Eingangsvektoren und einer Quantisierung nach der Multiplikation ausgegangen. Diese Werte werden dann addiert, wobei keine weitere Quantisierung stattfindet.

In Abbildung 5a ist das SNR aufgetragen über der Länge des Skalarprodukts n. Dabei wurden die Zufallsvektoren vor der Multiplikation auf 12 Bit gerundet. Nach der Multiplikation wurde auf BA (BA von 11 bis 15) Nachkommastellen abgeschnitten. Man erkennt, dass das SNR abhängig ist von der Länge n des Skalarprodukts. Werden dieselben Berechnungen für Runden durchgeführt, so ist das SNR unabhängig von n (siehe Abbildung 5b). Der Grund für die Abhängigkeit des SNR von n ist der Mittelwert des Quantisierungsfehlers, der beim Abschneiden ungleich Null ist. Dieser Mittelwert des Fehlers addiert sich auf und bewirkt eine Verschlechterung des SNR.

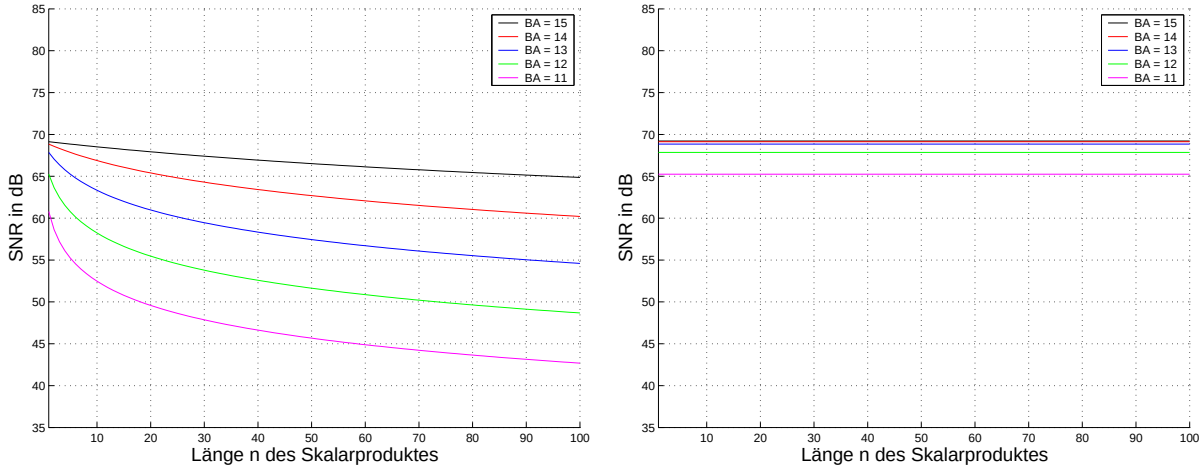

Abbildung 5. SNR des Skalarprodukts **(a)** beim Abschneiden **(b)** beim Runden.

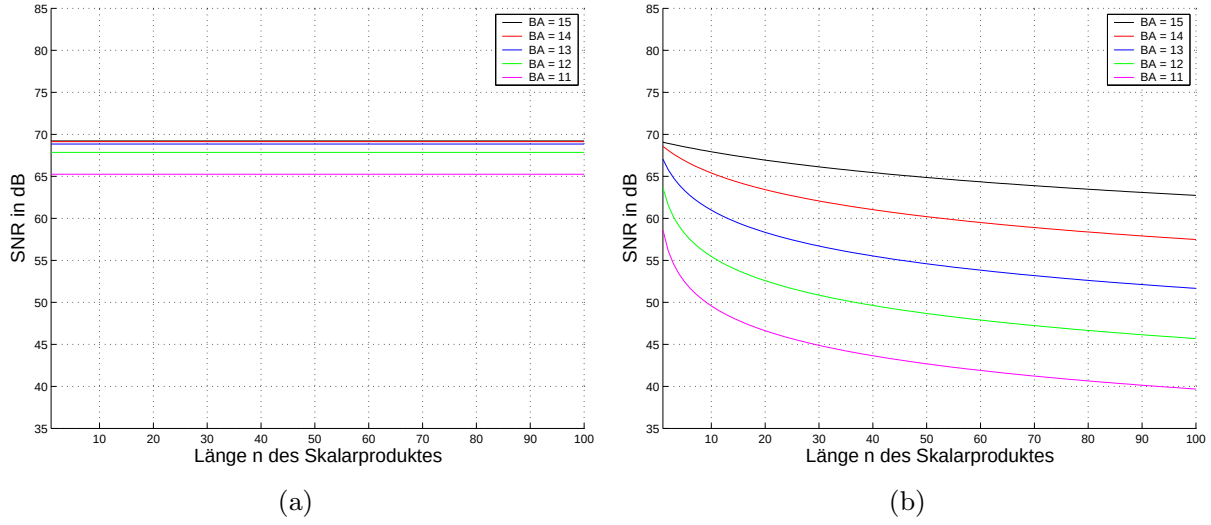

Abbildung 6. SNR des komplexen Skalarprodukts **(a)** Realteil, **(b)** Imaginärteil.

Nun soll die Auswirkungen der Quantisierung beim komplexen Skalarprodukt betrachtet werden. In Abbildung 6 ist das SNR für ein komplexes Skalarprodukt bei Verwendung von Abschneiden für die Quantisierung nach der Multiplikation aufgetragen über der Länge n. Dabei ist in Abbildung 6a das SNR des Realteils und in Abbildung 6b das SNR des Imaginärteils dargestellt. Wie man sieht, ist beim Realteil das SNR unabhängig von der Länge des Skalarproduktes, obwohl abgeschnitten wurde. Das SNR des Imaginärteils dagegen ist abhängig von der Länge des Skalarprodukts. Der Grund dafür wird ersichtlich, wenn man die Berechnung des Realteils und Imaginärteils beim Skalarprodukt genauer betrachtet ($[\ ]_Q$ steht für die Quantisierung nach der Multiplikation):

$$\Re = \sum_n [x_\Re \cdot y_\Re]_Q - \sum_n [x_\Im \cdot y_\Im]_Q \tag{12}$$

$$\Im = \sum_n [x_\Re \cdot y_\Im]_Q + \sum_n [x_\Im \cdot y_\Re]_Q \tag{13}$$

Für den Realteil werden zwei Summen subtrahiert, deshalb subtrahieren sich auch die Mittelwerte der Fehler und der resultierende Mittelwert ist Null. Bei der Imaginärteilberechnung addieren sich die Mittelwerte und deshalb ist der resultierende Mittelwert ungleich Null und verschlechtert das SNR. Um dies zu umgehen, gibt es mehrere Möglichkeiten. Eine Möglichkeit ist, anstatt abzuschneiden nach der Multiplikation zu runden. Dazu müsste nach jeder Multiplikation gerundet werden (insgesamt $2{\cdot}n$ mal). Eine andere Möglichkeit wäre, einen Faktor in einer Summe zu negieren und anschließend die Summe zu subtrahieren anstatt zu addieren.

$$\Im = \sum_n [x_\Re \cdot y_\Im]_Q - \sum_n [-x_\Im \cdot y_\Re]_Q \tag{14}$$

Dafür sind insgesamt n Negationen erforderlich und man erreicht damit dasselbe SNR wie für den Realteil bzw. wie für Runden. Die für manche Realisierungen einfachste Methode ist das Addieren eines Offsets, der genau dem negativen Mittelwert des Fehlers des Imaginärteils entspricht.

$$\Im = \sum_n [x_\Re \cdot y_\Im]_Q + \sum_n [x_\Im \cdot y_\Re]_Q + 2 \cdot n \cdot (-\mu_e) \qquad (15)$$

Diese Methode kann mit nur einer zusätzlichen Addition ausgeführt werden und erreicht dasselbe SNR wie Runden. Beim Runden würde nach jeder Multiplikation der Wert $2^{-\mathrm{B2}-1}$ (dies entspricht einer Eins an der ersten Stellen hinter der B2ten Stelle) addiert und danach auf B2 Stellen abgeschnitten. Die Summe dieser Einzeladditionen entspricht ungefähr dem Wert, der in Gleichung (15) addiert wird. Allgemein kann gesagt werden, dass der Fehler von Runden und Abschneiden sich lediglich im Mittelwert unterscheidet. Beim Runden ist der Mittelwert Null, beim Abschneiden nicht. Werden nach der Quantisierung nur lineare Berechnungen wie in Gleichung (15) durchgeführt, so kann der Mittelwert des Fehlers am Ausgang des Systems ausgeglichen werden (sofern das System bekannt ist). Werden jedoch nichtlineare Berechnungen durchgeführt, so ist es notwendig, den Mittelwert zuvor durch die Addition des negativen Mittelwertes auszugleichen, da sich der statistische Mittelwert durch die nichtlineare Operationen sich auf die Varianz auswirken kann und nicht mehr neutralisiert werden kann.

5 Zusammenfassung

In dem vorliegenden Beitrag wurden die Auswirkungen von Quantisierung vor und nach der Multiplikation von zwei Zufallszahlen untersucht. Weiter wurde das SNR für das reelle Skalarprodukt von zwei Zufallsvektoren für Runden und Abschneiden berechnet. Für den Imaginärteil des komplexen Skalarprodukts ergab sich bei Abschneiden nach den Multiplikationen eine Verschlechterung des SNR in Abhängigkeit von der Länge. Der Grund hierfür ist hauptsächlich der Mittelwert des Fehlers, der ungleich Null ist. Um diesen Effekt auszugleichen und dasselbe SNR wie für Runden, allerdings mit weniger HW Aufwand zu erhalten, wurden verschiedene Möglichkeiten vorgestellt.

Literatur

Beichelt, F.: Stochastik für Ingenieure, B. G. Teubner, Stuttgart, 1995.

Constantinides, G. A., Cheung, P. Y. K., and Luk, W.: Truncation noise in fixed-point SFGs, IEE Electronics Letters, 35, 2012–2014, 1999.

Oppenheim, A. V. and Schafer, R. W.: Zeitdiskrete Signalverarbeitung, vol. 3, R. Oldenburg Verlag, Rosenheimerstraße 145, D-71671 München, 1999.

Shanmugan, K. S. and Breipohl, A. B.: Random Singals; Detection, Estimation and Data Analysis, John Wiley & Sons, New York, 1990.