Adv. Radio Sci., 7, 73–81, 2009
www.adv-radio-sci.net/7/73/2009/

Advances in Radio Science

Analyse und Untersuchung der Quantisierungseffekte beim Goertzel-Filter

A. Tchegho[1], H. Gräb[1], U. Schlichtmann[1], H. Mattes[2], and S. Sattler[2]

[1]Lehrstuhl für Entwurfsautomatisierung, Fakultät für Elektrotechnik und Informationstechnik, Technische Universität München, Germany
[2]Infineon Technologies, Neubiberg, Germany

Zusammenfassung. Digitale Filter können in praktischen Anwendungen nur mit endlicher Wortbreite realisiert werden. Deshalb müssen Quantisierungseffekte verstanden werden, um die Eigenschaften und die Performanz der digitalen Filter geeignet einstellen zu können. Die am stärksten beeinflussten Aspekte sind neben Frequenzgang, die Stabilität und das Signal-Rausch-Verhältnis. Zu geringe Wortbreiten führen zu drastischen Veränderungen der Filtereigenschaften. Eine zu konservativ gewählte Wortbreite hingegen, erhöht die Größe und somit die Anzahl der Operationen des Filters unnötigerweise. Ziel dieses Beitrages ist es, für eine bestimmte Filterklasse optimale Wortbreiten zu finden, welche sowohl die Anzahl der Berechnungen und damit Operationen minimieren als auch sicherstellen, dass vorgegebene Toleranzbereiche für spezifizierte Filtereigenschaften eingehalten werden können. Die Untersuchungen werden an einem Goertzel-Filter durchgeführt, welches aufgrund seiner effizienten Eigenschaften besonders für die spektrale Bewertung von Mixed-Signal Schaltungen und Systemen geeignet ist.

1 Einleitung

Die stetig, zunehmende Funktionalität von integrierten Schaltungen und Systemen führt heute bereits zu einem Komplexitätsgrad, welcher ohne zusätzliche Maßnahmen in Software, Firmware oder Hardware nicht mehr zuverlässig gemeistert werden kann. Zukünftige Anforderungen sind vermehrt im Bereich (Vernay, 2008) Sicherheit und Gewährleistung von zuverlässigen Produkten zu finden. Applikationsspezifische Methoden und Verfahren spielen hier die große Rolle bei der Problemlösung (De Micheli, 2008). Sie können maßgeschneidert oder jeweils an die wechselnden Randbedingungen angepasst werden.

Insbesondere für die Diagnose und Analyse von Systemverhalten über die Zeit und während des Betriebes ist es wichtig, solche geeigneten Verfahren zukünftig auszuwählen, welche in der Lage sind, diese anstehende Aufgaben ressourcensparend und schritthaltend durchführen zu können. In der digitalen Signalverarbeitung haben solche Transformationen ein besonderes Gewicht, mit denen die „zyklische“ Faltungsumme (Oppenheim et al., 1999) schnell und effizient berechnet werden kann. Eine bekannte Transformation, bei der die Anzahl der Werte im Orginal- und Bildbereich gleich ist, ist die Diskrete Fourier Transformation (DFT). Jedem N-Tupel aus dem Eingangsbereich kann dort eindeutig ein N-Tupel aus dem Bildbereich zugeordnet werden (Agrawal and Burrus, 1974). Die Lösung eines Problems im Bildbereich ist sinnvoll, wenn der Aufwand für die Transformation geringer ist als für die Lösung im Orginalbereich.

Geeignete Implementierungen dazu berücksichtigen immer die Architektur des Ziel-Systems und können so optimal integriert werden. Programmierbare Schaltkreise (FPGA – Field Programmable Gate Array) sind hier auf dem Weg, die digitale Signalverarbeitung genauso zu revolutionieren, wie die programmierbaren digitalen Signalprozessoren (DSP – Digital Signal Processor) ca. 25 Jahre früher. Viele der an vorderster Front eingesetzten Algorithmen, z.B. die Schnelle Fouriertransformation (FFT – Fast Fourier Transform), die Finite Response Filter (FIR) oder die Infinite Response Filter (IIR) – bis dazumal in ASICs oder programmierbaren DSPs realisiert worden – werden heute durch FPGAs ersetzt. Moderne FPGA Familien stellen heute arithmetische Blöcke zur Verfügung, welche Berechnungen mit hoher Verarbeitungsgeschwindigkeit und mit geringem Flächenverbrauch und Kosten ermöglichen (Dipert, 2000).

Kann man im Zielsystem auf Mikroprozessoren oder programmierbare DSPs – sogenannte eingebettete Systeme – zurückgreifen, ist es wichtig, die Algorithmen optimal auf

Correspondence to: A. Tchegho
(aurelien.tchegho@tum.de)

Published by Copernicus Publications on behalf of the URSI Landesausschuss in der Bundesrepublik Deutschland e.V.

die jeweilige Hardware umzusetzen. Pipelining ist dabei die Implementierungstechnik, welche erlaubt, viele Instruktionen während der Ausführung zeitlich überlappend auszuführen (Patterson and Hennessy, 1998). Dies ist der Schlüssel, um digitale Signalverarbeitung heute schneller zu machen. Für eine Implementierung der Algorithmen in Logik trifft dies für die on-the-fly (schritthaltende) Auswertung ebenfalls zu. Werden Informationen aus dem Bildbereich (Spektralbereich) für die Bewertung von Signalen und Testpunkten benötigt, ist der jeweilige Ressourcenbedarf für die Berechnung und Abspeicherung der Spektrallinien ausschlaggebend. Werden für die Auswertung der Diagnoseantwort z.B. nur wenige Spektrallinien benötigt, liefert der Goertzel Algorithmus (Tchegho et al., 2008) die optimale Implementierung in Bezug auf Platzbedarf und Verarbeitungsgeschwindigkeit. Die Arbeit ist folgendermaßen gegliedert: In Abschnitt 2 wird der Goertzel-Algorithmus vorgestellt. Das Verfahren wird erklärt und die Grundstruktur eines Goertzel-Filters angegeben. Abschnitt 3 geht auf die Quantisierungsfehler des Verfahrens ein. In Abschnitt 4 werden die verschiedenen Fehlerquellen vorgestellt und voneinander getrennt untersucht. Ergebnisse der Untersuchung sind in verschiedenen Tabellen und Gleichungen abgelegt, die Grundlage für die optimale Dimensionierung von Goertzel-Filtern sind. Abschnitt 5 fasst die Ergebnisse der Arbeit noch einmal zusammen.

2 Der Goertzel-Algorithmus

Der Goertzel-Algorithmus (Goertzel, 1958) ist ein iteratives Verfahren, mit dem beliebige Punkte des Spektrums eines Signals einzeln berechnet werden können. Der Algorithmus wird meistens in der Nachrichtentechnik für die Dekodierung von DTMF (Dual-Tone Multi-Frequency) Signalen eingesetzt (Chen, 1996).

2.1 Theorie

Ausgangspunkt der Herleitung ist die Definition der DFT $X[k]$ einer endlichen Folge $x[n]$ (Länge N).

$$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn}, \qquad \text{mit } W_N = e^{-J\frac{2\pi}{N}}. \tag{1}$$

Die Periodizität der Folge W_N^{-kn} kann ausgenutzt werden:

$$W_N^{-kN} = e^{J\frac{2\pi}{N}kN} = e^{J2\pi k} = 1\,. \tag{2}$$

Da Gleichung (2) gilt, kann die rechte Seite von Gleichung (1) mit W_N^{-kN} multipliziert werden, ohne die Gleichung zu beeinflussen.

$$\begin{aligned} X[k] &= W_N^{-kN} \sum_{n=0}^{N-1} x[n] W_N^{kn} \\ &= \sum_{n=0}^{N-1} x[n] W_N^{-k(N-n)} \end{aligned} \tag{3}$$

Als das zu erwartende Ergebnis definiert man die Folge $y_k[n]$, welche die Faltung zweier Folgen $f[n]$ und $g_k[n]$ darstellt:

$$y_k[n] = f[n] * g_k[n] = \sum_{m=-\infty}^{\infty} f[m] g_k[n-m] \tag{4}$$

mit

$$f[n] = \begin{cases} x[n] & 0 \le n \le N-1 \\ 0 & \text{sonst} \end{cases} \tag{5}$$

und

$$g_k[n] = W_N^{-kn} \qquad \forall n. \tag{6}$$

Daraus folgt:

$$y_k[n] = \sum_{m=0}^{N-1} x[m] W_N^{-k(n-m)}. \tag{7}$$

Aus Gleichungen (3) und (7) und der Tatsache, dass $f[n]$ eine endliche kausale Folge ist, folgt:

$$X[k] = y_k[n]\Big|_{n=N}. \tag{8}$$

$y_k[n]$ ist also die Antwort eines zeitdiskreten Systems (mit der Impulsantwort W_N^{-kn}) auf eine endliche Eingangsfolge $x[n]$. Der N-te Wert der Folge $y_k[n]$ ist gleich dem DFT-Wert $X[k]$.

Die übertragungsfunktion $H_k(z)$ des Systems findet man durch Transformation der Impulsantwort in den z-Bereich:

$$H_k(z) = \frac{1 - W_N^k z^{-1}}{1 - 2\cos(2\pi k/N) z^{-1} + z^{-2}} \tag{9}$$

Diese Übertragungsfunktion ist komplex und lässt sich in einem reellen und einem imaginären Teil teilen.

$$H_R(z) = \frac{1 - \cos(2\pi k/N) z^{-1}}{1 - 2\cos(2\pi k/N) z^{-1} + z^{-2}} \tag{10}$$

$$H_I(z) = \frac{\sin(2\pi k/N) z^{-1}}{1 - 2\cos(2\pi k/N) z^{-1} + z^{-2}} \tag{11}$$

Abbildung 1 zeigt das Pol-Nullstellen-Diagramm des linearen zeitdiskreten Systems mit der der Übertragungsfunktion $H_k(z)$. Das System besitzt ein konjugiertes komplexes Polpaar bei der Frequenz $\pm\omega_k$ (ω_k=$2\pi k/N$) und eine komplexe Nullstelle bei der Frequenz $-\omega_k$. Pole und Nullstelle liegen

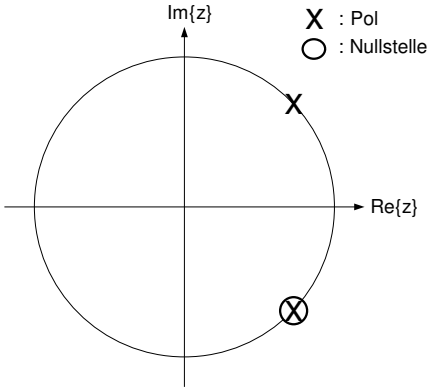

Abb. 1. Pol-Nullstellen-Diagramm des Systems $H_k(z)$ ($\omega_k = \pi/8$).

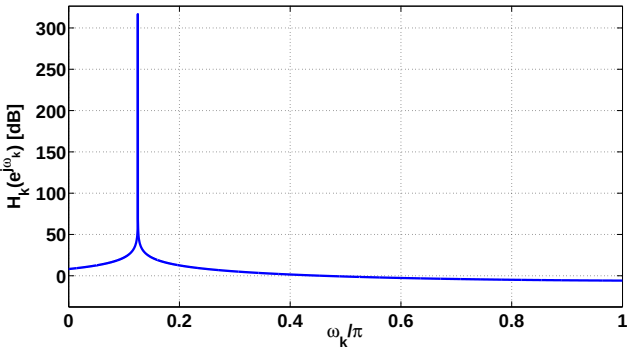

Abb. 2. Frequenzgang des Systems $H_k(z)$ ($\omega_k = \pi/8$).

auf dem Einheitskreis. Anzumerken sei hier, dass die Einführung des redundanten Pol-Nullstellen-Paars bei der Frequenz $-\omega_k$ auf eine ökonomische Hardware-Realisierung des Filters führt. Der Frequenzgang des Systems in Abhängigkeit des Frequenzindexes k wird in Abb. 2 gezeigt.

Der Signalflussgraph in Abb. 3 entspricht dem System gemäß Gleichung (9). Dies ist die Struktur eines digitalen rekursiven IIR(Infinite Impulse Response)-Filters zweiter Ordnung.

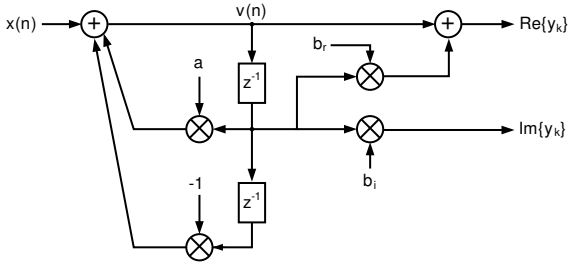

Abb. 3. Signalflussgraph des Goertzel-Filters mit $a = 2\cos(\omega_k)$, $b_r = -\cos(\omega_k)$, $b_i = \sin(\omega_k)$.

3 Quantisierungsfehler

Der Goertzel-Algorithmus stellt ein lineares zeitinvariantes zeitdiskretes System dar und wird durch ein digitales IIR-Filter zweiter Ordnung implementiert. Die Performanz solcher digitalen Filter wird durch die numerische Quantisierung stark beeinflusst. Um eine optimale Dimensionierung des Filters zu erzielen, müssen die Quantisierungsfehler und deren Auswirkungen verstanden und analysiert werden.

3.1 Zahlendarstellung

Bei der Implementierung digitaler Filter werden Signale und Koeffizienten in einem digitalen Zahlensystem dargestellt, das stets nur eine endliche Wortbreite hat. Das Zweierkomplement-Format wird dabei am häufigsten genutzt. Eine reelle Zahl x wird mit einer endlichen Wortbreite ([B+1]-Bit) in der Zweierkomplement-Form dargestellt als

$$\hat{x} = Q_B[x] = X_m \left(-b_0 + \sum_{i=1}^{B} b_i 2^{-i} \right) = X_m \hat{x}_B, \qquad (12)$$

wobei X_m ein beliebiger Skalierungsfaktor ist und die Werte der b_i entweder 0 oder 1 sind. Die Größe b_0 wird als Vorzeichenbit bezeichnet. Die reelle Zahl $x = -5_{10}$ wird z.B. im Zweierkomplement (4-Bit) als

$$\begin{aligned} \hat{x} &= Q_3[-5_{10}] \\ &= 2^3 \cdot \left(-1 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} \right) \\ &= 2^3 \cdot (1 \bullet 011)_2 \\ &= X_m \hat{x}_3 \end{aligned}$$

dargestellt.

Die kleinste Differenz zwischen den Zahlen ist

$$\Delta = X_m 2^{-B}. \qquad (13)$$

Der Quantisierungsfehler ist definiert als

$$e = \hat{x} - x \quad . \qquad (14)$$

Die quantisierten Zahlen liegen im Bereich $-X_m \leq \hat{x} < X_m$. Der gebrochene Teil von $\hat{x}$ kann durch die Notation

$$\hat{x}_B = b_0 \bullet b_1 b_2 b_3 \cdots b_B \qquad (15)$$

dargestellt werden, wobei $\bullet$ das Binärkomma darstellt. X_m ist ein Skalierungsfaktor, der die Darstellung von Zahlen erlaubt, deren Betrag größer als Eins ist. Ein Wert $X_m = 2^0$ impliziert, dass das Binärkomma zwischen b_0 und b_1 des Binärwortes in Gleichung (15) liegt. Man spricht vom sogenannten $1QB$-Format (1 Stelle vor und B Stellen nach dem Binärkomma). Entsprechend impliziert ein Wert $X_m = 2^2$, dass das Binärkomma tatsächlich zwischen b_2 und b_3 liegt ($3Q[B-2]$-Format).

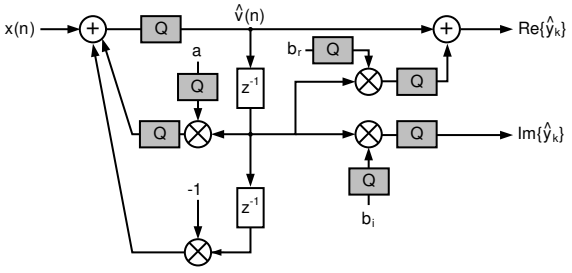

Abb. 4. Goertzel-Filter mit Fehlerquellen.

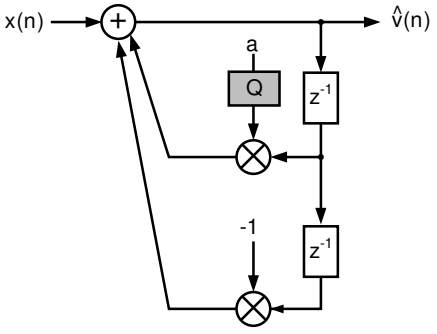

Abb. 5. Rekursiver Teil des Goertzel-Filters.

3.2 Fehlerquellen

Abbildung 4 zeigt ein Modell des Goertzel-Filters mit verschiedenen Fehlerquellen (Quantisierer Q). Das Eingangssignal liegt ebenfalls in quantisierter Form vor. Die Systemkoeffizienten (a, b_r, b_i) werden mit einer endlichen Wortbreite ([B+1]-Bit) dargestellt. Die Register sind damit [B+1]-Bit breit. Wird die verzögerte [B+1]-Bit Variable $\hat{v}[n-1]$ mit dem [B+1]-Bit Koeffizient $\hat{a}$ multipliziert, entsteht ein [2(B+1)]-Bit Produkt. Bei der Verwendung eines [B+1]-Bit-Addierers muss dieses Produkt auf [B+1]-Bit quantisiert werden. Entsteht bei der Addition zweier großer Zahlen eine Zahl, welche nicht mehr mit der gewählten Anzahl von Bits darstellbar ist, kommt es zu einem Überlauf. Infolge der Quantisierer (Q-Blöcke) und der Möglichkeiten eines Überlaufes an den Addierern ist das System nichtlinear.

Die Auswirkungen der Quantisierer können nicht genau beschrieben werden, da sie von der jeweiligen Eingangsfolgen abhängen und diese meistens unbekannt sind. Eine gute Abschätzung des Fehlers erreicht man jedoch, indem man die stochastische Natur des Quantisierungsfehlers heranzieht und die Auswirkungen der Quantisierung von Systemkoeffizienten getrennt von der endlichen Arithmetik untersucht.

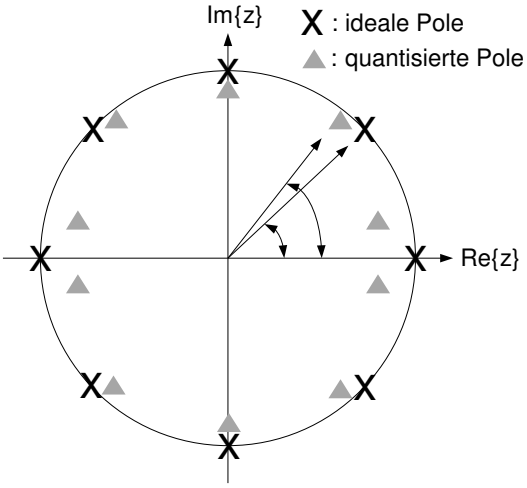

Abb. 6. Alle mögliche Pollagen und dazugehörige quantisierte Pole.

4 Ergebnisse

4.1 Auswirkungen der Quantisierung von Systemkoeffizienten

Aufgrund der Quantisierung können die Pole bzw. Nullstelle des Systems nur begrenzte Positionen in der z-Ebene einnehmen. Es kommt zu einer Verschiebung der Pole bzw. Nullstellen von deren idealen Position zur quantisierten Position. Dadurch ändert sich die Übertragungsfunktion.

$$\hat{H}_k(z) = \frac{1 - \mathrm{Q}[W_N^k]z^{-1}}{1 - \mathrm{Q}[2\cos(2\pi k/N)]z^{-1} + z^{-2}} \tag{16}$$

Beim Goertzel-Filter legen die Pollagen die Frequenz der zu berechnenden spektralen Komponente fest. Verschiebungen der Pollagen führen also zu Abweichungen in der gewünschten Frequenz. Im Folgenden wird nur der rekursive Teil (Abb. 5) des Goertzel-Filters, welcher für die Realisierung der Pole zuständig ist, betrachtet.

Abbildung 6 zeigt das Pol-Nullstellendiagramm des Systems für ein Beispiel N=8. Eingetragen sind die idealen Pollagen und die quantisierten Pollagen. Die quantisierten Pole liegen nicht mehr auf dem Einheitskreis. Dies führt zu einem Fehler bei der Amplitude der zu berechnenden spektralen Komponente. Viel wichtiger ist aber die Tatsache, dass die quantisierten Pole nicht mehr bei der Frequenz ω_k liegen, sondern bei der Frequenz

$$\hat{\omega}_k = \omega_k + \Delta\omega_k \,. \tag{17}$$

Abbildung 7 zeigt den Frequenzgang eines Goertzel-Filters mit 6-Bit quantisierten Koeffizienten. Die gestrichelte Linie stellt den Frequenzgang bei idealen Koeffizienten dar. Pole, die in der Nähe von ω_k=0 oder ω_k liegen, können sich mehr verschieben als die in der Nähe von ω_k=$\pi/2$.

Tabelle 1 gibt den relativen Frequenzfehler ($\Delta\omega_k/\omega_k$) in Abhängigkeit der Wortbreite der Koeffizienten wieder. Als

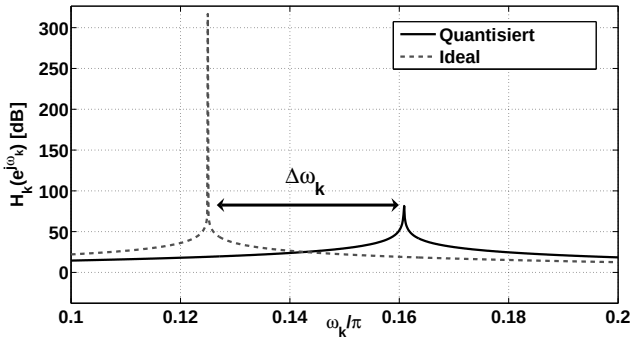

Abb. 7. Auswirkung der Koeffizientenquantisierung auf den Frequenzgang des Systems.

Parameter wird die Variable N verwendet. In der linken Tabellenhälfte sind die relativen Frequenzfehler ($\Delta\omega_k/\omega_k$) aus dem Frequenzbereich ω_k (k=0 . . . $N-1$) eingetragen. In der rechten Tabellenhälfte wird nur der reduzierte Frequenzbereich ($20\%N \leq k \leq 40\%N$) betrachtet. Von den berechneten Werten werden nur die maximalen Werten in die Tabelle eingetragen. Wie erwartet sinkt der maximale Fehler bei steigender Wortbreite. Je mehr Bits zur Verfügung stehen, desto geringer wird der Abstand zwischen quantisierten und idealen Pollagen. Bei konstanter Wortbreite ändert sich der Fehler in Abhängigkeit vom Parameter N. Je größer N wird, desto feiner wird die Diskretisierung der Frequenzachse und desto größer wird der maximale relative Fehler.

Der maximale relative Fehler wird durch Pole bestimmt, welche in der Nähe der reellen z-Achse liegen. Wird nur einen Frequenzbereich betrachtet, welcher weit entfernt von der reellen Achse (z.B. $20\%N \leq k \leq 40\%N$) liegt, stellt man fest, dass der maximale relative Fehler schon bei wenigen Bits extrem klein ist. Für N=512 und bei 6-Bit Wortbreite ist der relative Fehler bereits unterhalb 10%. Tabelle 1 dient als Grundlage für die Auswahl der Wortbreite (Bits) bei gegebener Randbedingung (Frequenzauflösung $\Delta\omega_k$).

4.2 Auswirkungen der endlichen Arithmetik

In diesem Abschnitt werden die Auswirkungen der endlichen Arithmetik untersucht, welche durch die Operationen (Multiplikation, Akkumulation) während der Filterberechnung auftreten. Diese können die Eigenschaften des Goertzel-Filters erheblich beeinflussen. Hierzu gehören insbesondere die Fehler durch Überläufe und durch Runden. Überläufe stellen die kritischere Fehlerursache dar, da sich ein System dadurch nichtlinear verhält, und große Fehler entstehen können, welche sogar das Filter unzuverlässig machen können. Es steht ja im Betrieb nicht wirklich fest, welches Signal denn tatsächlich analysiert oder berechnet wird.

Tabelle 1. Maximaler Frequenzfehler $\Delta\omega_k/\omega_k$ (in %) für verschiedene Werte von N in Abhängigkeit der Wortbreite (Bits) ($^{(*)}$: k=0 . . . $N-1$; $^{(+)}$: k=20%N . . . 40%N).

Bits	max $(\Delta\omega_k/\omega_k)^{(*)}$			max $(\Delta\omega_k/\omega_k)^{(+)}$		
	N=2^5	N=2^7	N=2^9	N=2^5	N=2^7	N=2^9
6	27.6	100	100	1.8	6.5	7.6
8	9.8	100	100	0.6	2.4	3.6
10	0.8	27.3	100	0.1	0.4	1.6
12	0.4	9.9	100	0.02	0.1	0.6
14	0.12	0.6	27.3	0	0.01	0.1
16	0.02	0.6	9.9	0	0	0.03

4.2.1 Überläufe

Bei der Addition zweier Zahlen im Zweikomplement kann es zu einem Überlauf kommen. Die resultierende Zahl lässt sich nicht mehr mit dem gewählten Format darstellen. Die Addition wird im Falle des Überlaufs zu einer nichtlinearen Operation. Abhilfe schafft hier eine geeignete Skalierung des Eingangssignals des Filters, so dass es nicht zu Überläufen kommen kann. Da das Eingangssignal durch die Skalierung verkleinert wird, verschlechtert sich das Signal/Rauschverhältnis des Ausgangssignals. Eine bessere Abhilfemaßnahme beim Goertzel-Filter ist die geeignete Formatanpassung an der jeweiligen Messaufgabe. Ein Kompromiss zwischen zwei widersprüchliche Forderungen muss gefunden werden:

- Überlauf soll vermieden werden.
- Genauigkeit (Anzahl der Nachkommastellen) soll maximal sein.

Um den Überlauf beim Goertzel-Filter zu vermeiden, soll eine bestimmte Anzahl V-Binärstellen vor dem Komma vorhanden sein. Bei einer Gesamtwortbreite von W-Bits führt dies zu dem Format $[VQ(W-V)]$. Dadurch erreicht man zwar eine höhere Dynamik (der Bereich der darstellbaren Zahlen wird größer), allerdings auf Kosten der Genauigkeit (sehr großer Abstand zwischen zwei aufeinanderfolgenden Zahlen). Die Wahl des optimalen Formats bzw. der Variable V wird gezielt an der Messaufgabe angepasst. Gesucht wird die maximal zulässige Amplitude $A_{\max,\mathrm{dB}}$, welche eine spektrale Komponente noch haben darf, damit bei der Filterberechnung Überläufe nicht auftreten. Dazu werden verschiedene Simulationen durchgeführt. Die Ergebnisse sind in der Tabelle 2 eingetragen. Die Amplituden (in dB) werden als Funktion der Anzahl der Akkumulationen N (Spalten) und der Anzahl der Binärstellen V vor dem Komma (Zeilen) angegeben.

Tabelle 2. Maximal zulässige Amplitude (in dB) als Funktion der Anzahl der Akkumulationen N und der Anzahl der Binärstellen vor dem Komma V.

	$N=2^5$	$N=2^6$	$N=2^7$	$N=2^8$	$N=2^9$
$V=1$	−40	−52	−64	−76	−89
$V=2$	−34	−46	−57	−70	−82
$V=3$	−28	−40	−51	−64	−77
$V=4$	−22	−34	−45	−58	−71
$V=5$	−16	−28	−39	−52	−65
$V=6$	−10	−22	−33	−45	−59

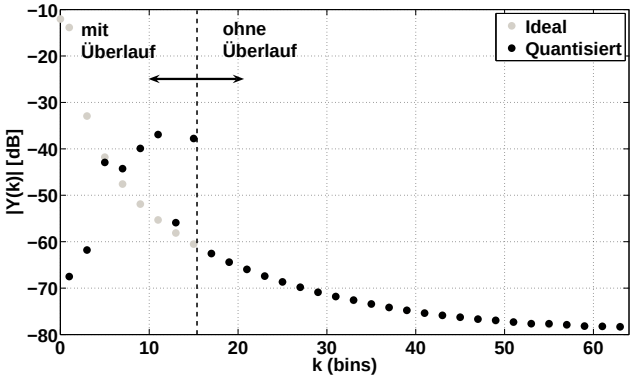

Abb. 8. Leistungsdichtespektrum in Abhängigkeit des Frequenzindexes k.

Aus der Tabelle 2 wird durch Interpolation folgende Gleichung ermittelt:

$$A_{\max,\mathrm{dB}} = f(N,V) = 20\log\left(\frac{\left(\frac{4}{N}\right)^2}{\sqrt{2}}\frac{1}{2^{V-1}}\right) . \qquad (18)$$

Gleichung (18) ist eine sehr gute Abschätzung für die maximal zulässige Amplitude, die die spektrale Komponente haben darf, damit Überläufe nicht auftreten. Ein Beispiel ist in Abb. 8 angegeben. Dargestellt ist das Leistungsdichtespektrum $Y[k]$ eines Dreiecksignals (N=128), welches mit dem Goertzel-Filter berechnet wird. Das Spektrum wird einmal mit Gleitkommazahlen (ideal) und einmal mit Festkommazahlen (quantisiert) berechnet. Die Berechnung wird mit einer Wortbreite von 16 Bits durchgeführt. Als Format für die interne Berechnung wird das [1Q15]-Format (V=1) verwendet. Ab ca. $A_{\max,\mathrm{dB}}$=−64 dB treten Überläufe nicht auf. Sollen z.B. Komponenten mit maximaler Amplitude -40 dB bestimmt werden, dann wird das Format [5Q11] (V=5) gewählt.

4.2.2 Rundungsrauschen

Bei der Multiplikation im Zweierkomplement kommt es zwar nicht zu einem Überlauf, allerdings erhöht sich im Allgemeinen die Anzahl der benötigten Binärstellen. Das Produkt von zwei $[B+1]$-Bit Zahlen benötigt $[2B+1]$ Bits für seine exakte Darstellung. Wird das Produkt nicht mit der vollen Wortlänge weiterverarbeitet, so wird auch die Multiplikation zu einer nichtlinearen Operation. Spätestens nach der Addition wird eine Wortlängenverkürzung vorgenommen, wenn die Werte in den $[B+1]$-Bit Register abgespeichert werden müssen. Abbildung 9 zeigt das nichtlineare Modell des Goertzel-Filters mit Rundungsfehler (Darstellung durch Quantisierer Q). Die Multiplikation mit -1 und die anschliessende Addition können einfach durch eine Substraktion ersetzt werden und erzeugen somit kein Rundungsrauschen. Ein lineares Modell (siehe Abb. 10) erhält man, indem die Quantisierer Q durch lineare Rauschquelle (e_1, e_2, e_3) ersetzt werden. Das hier gezeigte lineare Rauschmodell gestattet uns, das im System erzeugte Rundungsrauschen durch stochastische Werte wie Mittelwert und Varianz zu charakterisieren. Folgende Annahmen müssen allerdings hierzu getroffen werden:

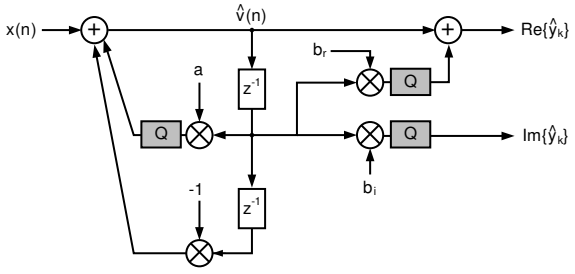

Abb. 9. Rundungsfehler beim Goertzel-Filter.

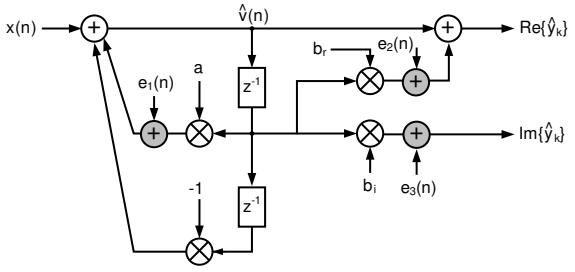

Abb. 10. Lineares Modell des Goertzel-Filters.

– Das Eingangssignal ist rein zufällig.

– Jede Rauschquelle stellt ein stationäres Zufallsprozess dar.

– Jede Rauschquelle ist unkorreliert mit dem Eingangssignal des jeweiligen Quantisierers.

– Das Quantisierungsrauschen der Rauschquelle ist über ein Quantisierungsintervall gleichverteilt.

Unter diesen Annahmen kann ein $[B+1]$-Bit Quantisierer durch eine lineare Rauschquelle mit der Varianz

$$\sigma_e^2 = \frac{2^{-2B}}{12} X_m^2 \tag{19}$$

ersetzt werden.

Wie bereits im Abschnitt 2 erwähnt, ist der N-te Wert der Ausgangsfolge des Goertzel-Filters gleich dem DFT-Wert. Es ist wichtig zu wissen, wie dieser Wert durch das im System erzeugte Rundungsrauschen beeinflusst wird. Wie groß ist also die Varianz des N-ten Wertes der Ausgangsfolge des Systems. Da das System linearisiert wurde, kann die Varianz der Ausgangsfolge über die Impulsantwort des Systems bestimmt werden. Die lineare Rauschquelle $e_1(n)$ stellt ein Störsignal dar, welches dem Eingangssignal $x(n)$ einfach überlagert wird. Dieses Signal wird also wie die Eingangsfolge $x(n)$ durch das System gefiltert. Die linearen Rauschquellen $e_2(n)$ und $e_3(n)$ werden ebenso der Ausgangsfolge überlagert. Die Varianz des N-ten Wertes der Ausgangsfolge (reeller und imaginärer Teil) wird also zu

$$\sigma_R^2 = \sigma_{e_2}^2 + \sigma_{e_1}^2 \sum_{n=0}^{N} h_R^2(n) \tag{20}$$

$$\sigma_I^2 = \sigma_{e_3}^2 + \sigma_{e_1}^2 \sum_{n=0}^{N} h_I^2(n) \quad . \tag{21}$$

Die Impulsantworten $h_R(n)$ und $h_I(n)$ erhält man durch inverse z-Transformation der übertragungsfunktionen (siehe Gleichungen 10 und 11).

$$h_R(n) = \cos(\omega_k n) \tag{22}$$

$$h_I(n) = \sin(\omega_k n) \quad . \tag{23}$$

Daraus folgt für den reellen Teil

$$\begin{aligned} \sum_{n=0}^{N} h_R^2(n) &= \sum_{n=0}^{N} \cos^2(\omega_k n) \\ &= \frac{1}{4} \sum_{n=0}^{N} \left(e^{\jmath\omega_k n} + e^{-\jmath\omega_k n}\right)^2 \\ &= \frac{1}{4} \sum_{n=0}^{N} \left(e^{2\jmath\omega_k n} + 2 + e^{-2\jmath\omega_k n}\right) \quad . \end{aligned} \tag{24}$$

Über die geometrische Reihe bekommt man

$$\sum_{n=0}^{N} e^{2\jmath\omega_k n} = \sum_{n=0}^{N} e^{-2\jmath\omega_k n} = 1 \quad . \tag{25}$$

Es folgt schließlich

$$\sum_{n=0}^{N} h_R^2(n) = \frac{N}{2} + 1 \tag{26}$$

Tabelle 3. Varianzen aus Gleichungen (28) und (29) in Abhängigkeit der Wortbreite B (Bits) und für N=128.

B	σ_R^2 [dB]	σ_I^2 [dB]
4	−16.63	−16.75
6	−28.72	−28.79
8	−40.76	−40.83
10	−52.8	−52.87
12	−64.84	−64.91
14	−76.88	−76.95

Ähnlich bekommt man für den imaginären Teil:

$$\sum_{n=0}^{N} h_I^2(n) = \frac{N}{2} \tag{27}$$

Die Varianz des N-ten Wertes der Ausgangsfolge ist damit

$$\begin{aligned} \sigma_R^2 &= \sigma_{e_2}^2 + \sigma_{e_1}^2 \left(\frac{N}{2} + 1\right) \\ &= \frac{2^{-2B}}{12} X_m^2 \left(\frac{N}{2} + 2\right) \end{aligned} \tag{28}$$

$$\sigma_I^2 = \frac{2^{-2B}}{12} X_m^2 \left(\frac{N}{2} + 1\right) \quad . \tag{29}$$

Gleichungen (28) und (29) stellen vereinfachte Abschätzungen des internen Rundungsrauschens des Filters dar und sollen zur Auswahl der internen Wortbreite beim Goertzel-Filter dienen. Tabelle 3 gibt die Werte der Varianzen für verschiedene Wortbreiten (B – Anzahl der Nachkommastellen) an.

4.3 Simulationen

Zur Herleitung des Rundungsrauschens beim Goertzel-Filter wurden verschiedenen Annahmen getroffen. Es wurde z.B. angenommen, dass alle Signale stationäre Zufallssignale (weisses Rauschen) sind und dass die Rauschquellen miteinander unkorreliert sind. Für die vorgesehene Anwendung als hochauflösender Spektralanalysator trifft dies beim Goertzel-Filter nicht zu. Hier werden deterministische Signale zugrunde gelegt. Zwei Fragen müssen beantwortet werden:

- Wie groß ist der aufgrund des Rundungsrauschens hervorgerufene Fehler, wenn deterministische Signale verwendet werden?
- Wie stark weicht der tatsächliche Fehler von dem mithilfe von Gleichung (28) und (29) abgeschätzten Wert ab?

Für die Simulation wird nun ein breitbandiges Dreiecksignal (Abb. 11) verwendet. Verschiedene Sequenzlängen werden jetzt eingestellt (N=32, 64, 128). Das Rundungsrauschen wird für verschiedene Wortbreiten (B) untersucht. Abbildung 12 zeigt die Varianz des Goertzel-Filterausgangs in

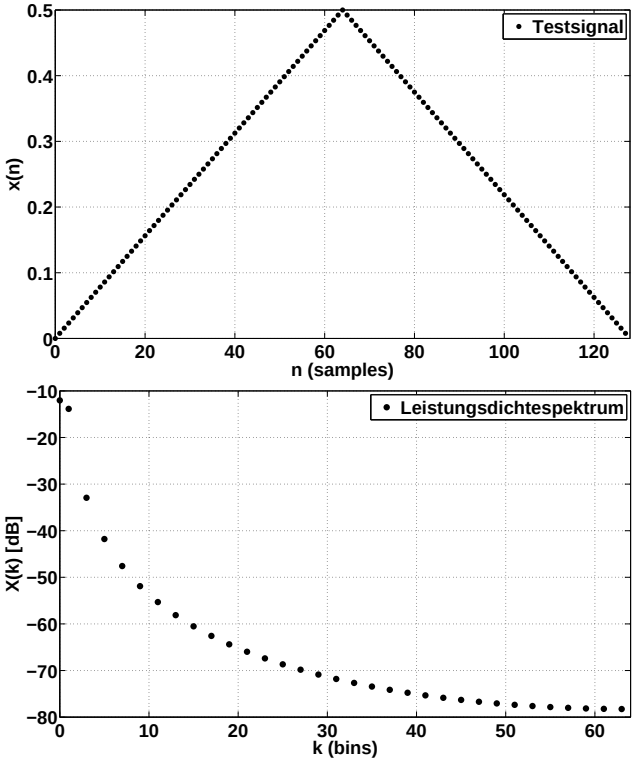

Abb. 11. Testsignal im **(a)** Zeit- und **(b)** Frequenzbereich für N=128.

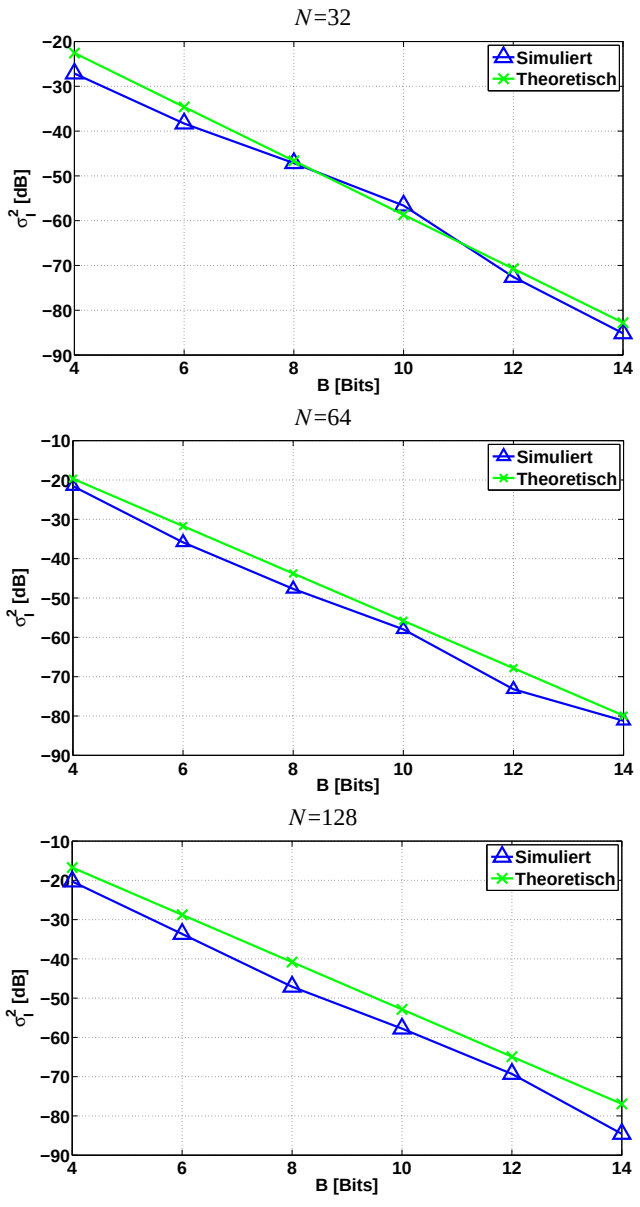

Abb. 12. Vergleich zwischen theoretischer und simulierter Varianz σ_I^2.

Abhängigkeit der Wortbreite. Das Spektrum des Testsignals (Dreiecksignal) weist nur imaginäre Komponenten auf. Deshalb wird auch nur die Varianz (σ_I^2) des imaginären Ausgangs (Im$\{y_k\}$) untersucht. Wie man aus den Abbildungen entnehmen kann, sinkt die Varianz des Rundungsrauschens mit steigender Wortbreite. Für alle Werte von N ist die simulierte Varianz im Allgemeinen kleiner als die theoretische Varianz. Weiterhin kann man eine gewisse Proportionalität zwischen der theoretischen und der simulierten Varianz erkennen. Die Abweichungen sind auf Korrelationen zurückzuführen, welche durch die deterministische Natur des Eingangssignals hervorgerufen werden. Bei dem hier verwendeten Testsignal sind Gleichungen (28) und (29) hinreichende Abschätzungen, welche als Grundlage für die Dimensionierung der internen Wortbreite des Filters herangezogen werden.

5 Zusammenfassung

Bei der Realisierung digitaler Filter mit Festkommazahlen treten Quantisierungsfehler auf. Die Fehlerquellen – Quantisierung der Systemkoeffizienten, Überläufe, Rundungsrauschen – werden am Beispiel eines Goertzel-Filters identifiziert und analysiert. Eine Tabelle wird aufgestellt, anhand derer man die optimale Wortbreite bei geforderter Frequenzauflösung auswählen kann. Es wird auch gezeigt, dass Überläufe – welche das Filter nichtlinear und unzuverlässig machen können – durch eine geeignete Formatanpassung an die jeweilige Aufgabe weitgehend vermieden werden. Weiterhin wird unter geeigneten Annahmen ein theoretisches Modell zur Abschätzung des im Filter erzeugten Rundungsrauschens hergeleitet. Das Modell wird anhand eines Dreiecksignals validiert. Damit steht einer Anwendung des Algorithmus im produktiven Umfeld nichts mehr im Wege. Der Goertzel-Algorithmus kann für die zuverlässige Analyse von spektraler Komponente verwendet werden.

Literatur

Agrawal, R. and Burrus, C.: Fast Convolution Using Fermat Number Transforms with Applications to Digital Filtering, IEEE Transactions on Acoustics, Speech and Signal Processing, 22, 87–97, 1974.

Chen, C.: Modified Goertzel Algorithm in DTMF Detection Using the TMS320C80, Texas Instruments, 1996.

De Micheli, G.: Designing Micro/Nano Systems for a Safer and Healthier Tomorrow, in: Design Automation and Test (DATE), pp. 1–1, 2008.

Dipert, B.: EDN's First Annual Programmable Logic Directory, EDN, pp. 54–84, 2000.

Goertzel, G.: An Algorithm for the Evaluation of Finite Trigonometric Series, The American Mathematical Monthly, 65, 34–35, 1958.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R.: Discrete-Time Signal Processing, Pearson Education, Inc., 2nd edn., 1999.

Patterson, D. A. and Hennessy, J. L.: Computer Organization and Design, Morgan Kaufmann Publishers, 2nd edn., 1998.

Tchegho, A., Mattes, H., and Sattler, S.: Optimaler Hochauflösender Spektralanalysator, Entwicklung von Analogschaltungen mit CAE-Methoden (ANALOG), pp. 183–188, 2008.

Vernay, D.: Perspective on Embedded Systems: Challenges, Solutions and Research Priorities, in: Design Automation and Test (DATE), pp. 2–2, 2008.