CAPITOLUL 5

EFECTELE LUNGIMII FINITE A CUVINTELOR ÎN FILTRAREA DIGITALĂ

5. 1. Introducere

Teoria filtrelor digitale s-a bazat pe presupunerea că atât semnalele, cât şi parametrii filtrelor pot avea orice valoare finită. În realitate, datorită limitării lungimilor cuvintelor din orice sistem digital, sunt permise numai valori discrete ale amplitudinii semnalelor, respectiv coeficienților. Luând în considerație aceste valori discrete în relațiile care caracterizează filtrele, vor rezulta ecuații neliniare, care, în general, nu vor putea fi riguros prelucrate.

Implementarea sistemelor discrete, fără a considera efectele lungimii finite a cuvintelor, inerente în orice implementare digitală, a condus la obținerea unor caracteristici liniare. De fapt, au fost analizate sisteme modelate liniar, dar ale căror realizări digitale sunt implicit neliniare. Această problemă reprezintă un dezavantaj major al filtrelor digitale şi, prin urmare, analiza efectelor lungimii finite a cuvintelor asupra performanțelor filtrelor constituie o etapă importantă în proiectarea filtrelor digitale.

În cazul filtrelor recursive, caracteristicile neliniare rezultate din operația de cuantizare din multiplicatoare, pot cauza un comportament oscilatoriu la ieşirea filtrelor, chiar şi în absența semnalului de intrare. Mai mult, în sumatoare poate apărea depăşirea aritmetică care produce, de asemenea, oscilații la ieşire.

În cazul calculatoarelor care lucrează cu lungimi mari ale cuvintelor (adică au un număr mare de biți disponibili pentru reprezentarea numerelor), efectele cuantizării pot fi nesemnificative. Acestea cresc cu descreşterea numărului de biți. Din acest motiv sunt necesare modele matematice care să permită estimarea efectelor lungimii finite a cuvintelor asupra performanțelor filtrelor. Un model simplu este

cel care se bazează pe presupunerea că erorile de cuantizare sunt mici în comparație cu nivelul semnalului sau al parametrului, adică este o cuantizare „fină” în care erorile pot fi tratate ca zgomot şi problema devine liniară [23].

Principalele tipuri de erori de cuantizare care apar în filtrarea digitală sunt:

1. Erori de cuantizare ale semnalului de intrare în conversia analog – digitală (A/D);
2. Erori rezultate din cuantizarea coeficienților filtrelor digitale;
3. Erori rezultate din rotunjirea produselor;
4. Depăşirea aritmetică;
5. Oscilații cu cicluri limită.

Dintre aceste tipuri de efecte, erorile de cuantizare ale semnalului de intrare au loc în afara filtrului, înaintea calculelor interne, restul efectelor sunt interne filtrului şi influențează metoda prin care sistemul va fi implementat.

De exemplu, pentru un filtru digital de ordinul întâi

$$y[n] = Ay[n-1] + x[n] \tag{5.1}$$

eroarea de tipul 1 se referă la cuantizarea intrării $x[n]$, eroarea de tipul 2 apare în reprezentarea parametrului A iar cea de tipul 3 apare la formarea produsului $Ay[n-1]$, necesar la fiecare iterație.

Elementul de bază dintr-un calculator numeric este circuitul cu două stări echiprobabile, căruia i se asociază o informație de 1 bit. *N* astfel de dispozitive pot fi cascadate pentru a forma un registru care conține *N* biți de informație. Implementarea unui filtru digital recursiv de ordinul întâi descris de ecuația (5.1) şi redată în figura 5.1, ilustrează cele mai importante operații ce trebuie efectuate.

Ieşirea anterioară $y[n-1]$ este stocată în registrul de ieşire sub forma unui număr pe *N* biți. Acesta este multiplicat cu numărul pe *N* biți care reprezintă coeficientul *A* care a fost stocat în registrul pentru coeficienți. Produsul $A\,y[n-1]$ (după rotunjire la *N* biți) este adunat la intrarea curentă $x[n]$ (de asemenea un număr pe *N* biți) pentru a forma ieşirea actuală $y[n]$ care este stocată pentru multiplicare cu *A* în iterația următoare. Întreaga procedură începe cu o valoare inițială $y[-1]$ stocată în registrul de ieşire. Aceasta poate fi sau nu, egală cu zero. Filtrele de ordin superior pot fi implementate într-un mod similar.

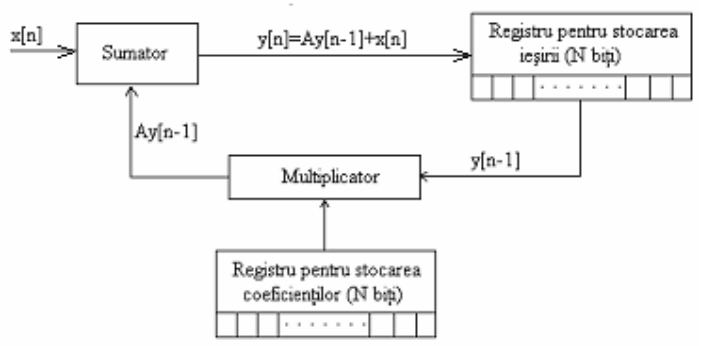

Figura 5.1. Implementarea unui filtru recursiv de ordinul întâi

Diferitele structuri de implementare ale unui sistem descris de ecuaţii cu diferenţe cu coeficienţi constanţi sunt echivalente dacă furnizează aceeaşi ieşire pentru o intrare dată, presupunând calculele interne ca fiind efectuate cu precizie infinită. Acestea nu sunt echivalente când sunt realizate cu precizie finită.

Trei factori importanţi contribuie la alegerea unei anumite realizări a filtrelor:

- complexitatea calculelor,
- necesarul de memorie,
- efectele lungimii finite a cuvintelor.

Efectul lungimii finite a cuvintelor reprezintă un factor important în implementarea sistemelor digitale de prelucrare a semnalelor şi trebuie luat în calcul la realizarea filtrelor digitale, deoarece limitarea numărului de biţi conduce la degradarea performanţelor filtrelor digitale. Înainte de a examina aceste efecte, se va prezenta o scurtă introducere în aritmetica digitală.

5.2. Reprezentarea numerelor

În procesarea digitală a semnalelor analogice, eşantioanele semnalului analogic sunt reprezentate în format digital. În principiu, procesul de conversie A/D implică eşantionarea semnalului analogic şi reprezentarea eşantioanelor ca secvenţe de biţi care definesc amplitudinea cuantizată a semnalului. Principala caracteristică a aritmeticii digitale constă în numărul limitat (de obicei fix) de biţi folosiţi în reprezentarea numerelor. Această constrângere are ca rezultat precizia finită a

calculelor, care conduce la erori şi efecte neliniare în comportamentul filtrelor digitale.

În cadrul reprezentării binare a numerelor reale sunt mai multe metode prin care un eşantion al unui semnal analogic poate fi reprezentat în format binar. Clasa reprezentărilor binare poate fi împărţită în reprezentările în virgulă fixă, virgulă mobilă şi virgulă mobilă cu blocuri.

5.2.1. Reprezentarea numerelor în virgulă fixă

Reprezentarea numerelor în virgulă fixă este generalizarea reprezentării zecimale, în care numerele din stânga virgulei reprezintă partea întreagă a numărului, iar cele din dreapta virgulei, partea fracţionară.

$$x = \left(b_{-a} \cdots b_{-1}\ b_0, \cdots b_b\right)_r = \sum_{i=-a}^{b} b_i r^{-i} \qquad 0 \le b_i \le (r-1) \qquad (5.2)$$

unde b_i reprezintă cifra, r – baza, a+1 – numărul de cifre ale părţii întregi şi b – numărul de cifre ale părţii fracţionare.

Datorită vitezei şi costului scăzut al părţii hard asociate, reprezentarea în virgulă fixă este deseori preferată în computere mai puţin performante şi în circuite dedicate care lucrează în timp real. Cea mai cunoscută reprezentare este cea pentru care r=2, în care numerele b_i se numesc *numere binare* sau *biţi* şi pot lua valorile {0,1}, obţinându-se *codul binar natural direct*. „Virgula binară" dintre b_0 şi b_1 nu există fizic în calculator. Circuitele logice ale acestuia sunt proiectate astfel încât calculele să aibă ca rezultat numere ce corespund poziţiei virgulei binare. Totuşi, în cele ce urmează, se va folosi virgula pentru a sublinia caracterul fracţionar al numărului reprezentat.

Folosind un format întreg pe n biţi (a=n-1, b=0), se pot reprezenta întregi fără semn cuprinşi în domeniul $0 \div (2^n-1)$. De obicei se foloseşte formatul fracţionar (a=0, b=n-1), cu virgula binară între b_0 şi b_1, care permite reprezentarea numerelor în domeniul $0 \div (1 - 2^{-n})$.

Indiferent dacă codul binar reprezintă o fracţie, un întreg, sau ambele, primul bit din stânga este numit cel mai semnificativ bit (most significant bit, MSB) iar bitul cel mai din dreapta, cel mai puţin semnificativ bit (least significant bit, LSB). În reprezentarea unei fracţii, MSB are o pondere de 2^{-1}=1/2 iar LSB are o pondere de 2^{-b}=$1/2^b$, unde b este numărul de biţi pe care este reprezentată fracţia. Ponderea 2^{-b}=$1/2^b$ desemnată de LSB este numită şi *rezoluţie*.

Orice întreg sau număr cu parte întreagă şi fracționară poate fi reprezentat în format fracționar prin factorizarea termenului r^a în relația (5.2). În această notație un cuvânt de cod de a+1 biți, cum ar fi 10011, corespunde numărului întreg

$$A = 1\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 0\cdot 2^3 + 1\cdot 2^4 = 1+2+16 = 19$$

Pe de altă parte, numărul 0,10011 reprezintă o fracție corespunzătoare numărului zecimal

$$B = 1\cdot 2^{-1} + 0\cdot 2^{-2} + 0\cdot 2^{-3} + 1\cdot 2^{-4} + 1\cdot 2^{-5} = \frac{1}{2} + \frac{1}{16} + \frac{1}{32} = \frac{19}{32}$$

Se observă că o deplasare a virgulei binare spre stânga cu n poziții corespunde unei împărțiri a numărului cu 2^n, iar o deplasare a virgulei binare spre dreapta cu n poziții corespunde unei înmulțiri a numărului cu 2^n.

Pentru a transforma un număr zecimal în corespondentul său binar, se procedează astfel: se divide în mod repetat numărul zecimal din stânga virgulei la 2, reținându-se restul. Acesta, scris în ordine inversă (de la dreapta spre stânga) este reprezentarea binară a părții întregi. Partea din dreapta virgulei se multiplică în mod repetat cu 2, înlăturând de fiecare dată partea zecimală şi reținând partea întreagă. Scriind aceasta în ordine normală, (de la stânga la dreapta), se obține reprezentarea binară a părții fracționare.

Exemplul 5.1.

Să se transforme numărul zecimal 627,625 în format binar.

Soluție.

Partea întreagă		Partea zecimală	
627 : 2 = 313	1	0.625 x 2 = 1.250	1
313 : 2 = 156	1	0.250 x 2 = 0.500	0
156 : 2 = 78	0	0.500 x 2 = 1.000	1
78 : 2 = 39	0	0.000 x 2 = 0.000	0
39 : 2 = 19	1		
19 : 2 = 9	1		
9 : 2 = 4	1		
4 : 2 = 2	0		
2 : 2 = 1	0		
1 : 2 = 0	1		

Prin urmare $(627{,}625)_{10} = (1001110011{,}101)_2$

Operațiile cu numere binare se execută similar celor zecimale.

1. Adunarea
   0 + 0 = 0
   0 + 1 = 1
   1 + 0 = 1
   1 + 1 = 0 se transportă 1
2. Scăderea
   0 – 0 = 0
   1 – 0 = 1
   0 – 1 = 1 se importă 1
   1 – 1 = 0
3. Multiplicarea
   0 x 0 = 0
   1 x 0 = 0
   0 x 1 = 0
   1 x 1 = 1
4. Împărțirea
   1 : 1 = 1
   0 : 1 = 0
   împărțirea la 0 nu este definită.

Aritmetica în virgulă fixă este potrivită atât pentru operații cu numere întregi, cât şi fracționare.

Dacă este necesară rotunjirea produsului a două numere, este mai bine a se limita reprezentarea în virgulă fixă a numerelor fracționare, decât a celor care au atât parte întreagă, cât şi fracționară, deoarece reducerea numărului de biți ai părții întregi ar cauza erori mari.

În conversia semnalelor analogice bipolare, este necesar un bit adițional pentru a purta informația de semn. De obicei cel mai semnificativ bit este rezervat semnului numărului, cu convenția ca zero să indice un număr pozitiv, iar unu, un număr negativ. Rezultatul este un *cod bipolar*. Există mai multe posibilități de reprezentare a codurilor bipolare binare, alegerea dintre acestea făcându-se în funcție de avantajele şi dezavantajele pe care le prezintă fiecare pentru aplicația respectivă. Patru metode sunt frecvent folosite pentru reprezentarea numerelor bipolare. În continuare se va considera că numerele sunt reprezentate pe $N=b+1$ biți, din care unul pentru semn.

Formatul mărime cu semn sau *semn – valoare* este cea mai simplă metodă pentru reprezentarea numerelor cu semn în format digital. Un zero în poziția MSB reprezintă un număr pozitiv, iar un unu în aceeaşi poziție

reprezintă un număr negativ. Restul de *b* biți reprezintă *modulul* sau *amplitudinea* numărului.

În cazul numerelor fracționare, reprezentarea mărime cu semn pentru un număr pozitiv $x \geq 0$ este de forma

$$(x)_{ms} = 0, b_1 b_2 ... b_b, \tag{5.3}$$

iar pentru numărul negativ $x_N = -x = -0, b_1 b_2 \cdots b_b$, de forma

$$(x_N)_{ms} = 1, b_1 b_2 ... b_b, \tag{5.4}$$

Aşa cum s-a precizat deja, virgula nu există fizic în reprezentarea numărului, dar, în cele ce urmează va fi utilizată pentru a specifica numerele fracționare. Se observă că în acest format zero are două reprezentări: 0,0…0 şi 1,00…0.

Valoarea zecimală a unui număr fracționar pozitiv este

$$(x)_{ms} = \sum_{i=1}^{b} b_i 2^{-i}, \tag{5.5}$$

iar a unui număr fracționar negativ este

$$(x_N)_{ms} = -\sum_{i=1}^{b} b_i 2^{-i}. \tag{5.6}$$

Modulul unui număr fracționar reprezentat în formatul mărime cu semn este dat de

$$|x| = |x_N| = \sum_{i=1}^{b} b_i 2^{-i}. \tag{5.7}$$

Reprezentarea în *complement față de unu* este identică celei în reprezentarea mărime cu semn pentru numere pozitive, dar diferă prin modul cum sunt formate numerele negative. În acest format, un număr negativ este obținut prin complementarea numărului pozitiv corespunzător.

În cazul formatului fracționar, numerele pozitive se reprezintă ca în relația (5.3), iar cele negative $x_N = -x = -0, b_1 b_2 \cdots b_b$ sub forma

$$(x_N)_{1C} = \overline{0}, \overline{b_1}\,\overline{b_2} ... \overline{b_b} = 1, \overline{b_1}\,\overline{b_2} ... \overline{b_b} \tag{5.8}$$

Plecând de la relația (5.8), reprezentarea în complement față de unu a unui număr negativ fracționar mai poate fi exprimată în forma

$$(x_N)_{1C} = 1 \times 2^0 + \sum_{i=1}^{b} (1 - b_i) 2^{-i} = 2 - 2^{-b} - |x| \tag{5.9}$$

Se observă ambiguitate în reprezentarea lui zero, ca 0,0…0 sau 1,1…1.

Modulul numărului negativ $b_0, b_1 b_2 \cdots b_b$ reprezentat în complement față de unu este

$$|x_N| = 1 - \sum_{i=1}^{b} b_i 2^{-i} - 2^{-b} \tag{5.10}$$

Valoarea zecimală a numărului negativ $b_0, b_1 b_2 \cdots b_b$ reprezentat în complement față de unu este

$$\left((x_N)_{1C}\right)_{10} = -1 + \sum_{i=1}^{b} b_i 2^{-i} + 2^{-b} \tag{5.11}$$

Spre exemplu, reprezentarea lui −3/8 este 1,100, care este complementul față de unu al lui 0,011 (3/8).

Reprezentarea în *complement față de doi* este identică cu formatul mărime cu semn în cazul numerelor pozitive. Prin urmare numerele pozitive sunt reprezentate cu un zero în poziția bitului de semn. Pentru a obține reprezentarea în complement față de doi a unui număr negativ, se scrie modulul acestuia în formatul mărime cu semn, se inversează biții acestei reprezentări şi se adună o unitate logică în poziția LSB.

Similar, un număr fracționar pozitiv se reprezintă sub forma (5.3), iar numărul fracționar negativ $x_N = -x = -0, b_1 b_2 \cdots b_b$, sub forma

$$(x_N)_{2c} = \overline{0}, \overline{b}_1 \overline{b}_2 \cdots \overline{b}_b + 0,0 \cdots 01 \tag{5.12}$$

Semnul "+" indică adunarea modulo 2 care ignoră bitul de transport, dacă acesta este prezent în MSB.

Plecând de la relația (5.12), reprezentarea în complement față de doi a unui număr fracționar negativ mai poate fi exprimată în forma

$$(x_N)_{2C} = 1 + \sum_{i=1}^{b} (1 - b_i) 2^{-i} + 2^{-b} = 2 - |x|, \tag{5.12'}$$

adică, un număr fracționar negativ este complementul față de doi al numărului pozitiv corespunzător, care se obține scăzând numărul pozitiv din 2, reprezentat în binar. De aici provine denumirea formatului.

Din (5.9) şi (5.12') rezultă

$$(x_N)_{2C} = (x_N)_{1C} + 2^{-b} \tag{5.13}$$

Valoarea zecimală a unui număr $b_0, b_1 b_2 \cdots b_b$ reprezentat în complement față de doi, este

$$(x_{2C})_{10} = -b_0 2^0 + \sum_{i=1}^{b} b_i 2^{-i} \tag{5.14}$$

unde $b_0 = 0$, pentru numere pozitive şi $b_0 = 1$, pentru numere negative.

Modulului numărului negativ reprezentat în complement față de doi este

$$|x_N| = 1 - \sum_{i=1}^{b} b_i 2^{-i} \tag{5.15}$$

De exemplu, reprezentarea în complement față de doi a numărului −3/8 se obține din complementarea lui 0,011 (3/8), rezultând 1,100, şi apoi adăugând 0,001. Rezultatul final este 1,101.

Codul binar deplasat sau o*ffsetul binar* este similar codului binar direct, obținându-se din acesta prin deplasarea în domeniul valorilor negative cu jumătate din întreaga scală. Cu *b*+1 biți se pot reprezenta 2^{b+1} numere. Pentru un cod bipolar există 2M numere, cu $M=2^b$, cuprinse în intervalul $-2^b \div (2^b - 1)$ pentru numere întregi şi în intervalul $-1 \div (1-2^{-b})$ pentru numere fracționare. În acest format cel mai mic număr negativ este reprezentat de un număr format din *b*+1 biți de zero iar cel mai mare număr pozitiv este format din *b*+1 biți de unu. În acest caz zero are o singură reprezentare şi, prin urmare, se evită ambiguitatea întâlnită la formatul mărime cu semn. Marele dezavantaj al acestei notații este dat de posibilele erori ce pot apărea la citirea MSB-ului, în loc de unu, zero sau invers, rezultând o eroare de amplitudine mare.

Dacă se compară formatul complement față de doi şi offsetul binar, se constată că ele diferă prin MSB si, prin urmare, este uşor a se trece de la o reprezentare la alta.

În Tabelul 5.1 sunt date codurile bipolare prezentate pentru reprezentarea numerelor întregi pe 4 biți, dintre care unul pentru semn.

TABEL 5.1 Coduri bipolare

Număr	*Formatul mărime cu semn*	*Ofset binar*	*Complement față de doi*	*Complement față de unu*
7	0111	1111	0111	0111
6	0110	1110	0110	0110
5	0101	1101	0101	0101
4	0100	1100	0100	0100
3	0011	1011	0011	0011
2	0010	1010	0010	0010
1	0001	1001	0001	0001
0	0000	1000	0000	0000

0	1000	1000	0000	1111
-1	1001	0111	1111	1110
-2	1010	0110	1110	1101
-3	1011	0101	1101	1100
-4	1100	0100	1100	1011
-5	1101	0011	1011	1010
-6	1110	0010	1010	1001
-7	1111	0001	1001	1000
-8	-	0000	-	-

În Tabelul 5.2 sunt date, comparativ, diferite reprezentări ale numerelor fracționare pentru o lungime de 3 biți a cuvintelor.

Tabelul 5. 2

Număr binar	**Echivalentul zecimal folosind reprezentarea**		
	Mărime şi semn	**Complement față de 1**	**Complement față de 2**
0,11	3 / 4	3 / 4	3 / 4
0,10	2 / 4	2 / 4	2 / 4
0,01	1 / 4	1 / 4	1 / 4
0,00	0	0	0
1,00	- 0	- 3 / 4	- 4 / 4 = - 1
1,01	-1 / 4	- 2 / 4	- 3 / 4
1,10	- 2 / 4	- 1 / 4	- 2 / 4
1,11	- 3 / 4	- 0	- 1 / 4

Din tabel se observă, aşa cum s-a mai specificat, că există două reprezentări pentru zero în format mărime cu semn şi complement fată de 1 şi nici o reprezentare pentru –1. Formatul complement față de 2 are o singură reprezentare pentru 0 şi poate reprezenta numere cuprinse între –1 şi $1-2^{-2}$ sau, în general, între –1 şi $1-2^{-(N-1)}$ pentru un registru de N biți. Reprezentarea în *complement față de 2* este adesea utilizată în implementarea filtrelor digitale datorită uşurinţei efectuării operațiilor de adunare şi scădere, caz în care descăzutul se adună cu complementul față de doi a scăzătorului.

Diferența dintre numărul maxim şi cel minim ce poate fi reprezentată se numeşte *domeniu dinamic.*

Exemplul 5.2.

Folosind reprezentarea în complement față de 2 pe 4 biți să se efectueze operațiile a) A - B şi b) B - A unde A = 0,250 şi B = 0,625

Soluție

a)	zecimal	complement față de 2
	0,250 -	0,010 +
	0,625	1,011
	-0,375	1,101 = - 0,375
b)	0,650 -	0,101 +
	0,250	1,110
	0,375	0,011 = 0,375

Se observă că în reprezentarea în complement față de 2 bitul de transport în poziția cea mai semnificativă este neglijat.

Adunarea şi scăderea în complement față de 1 sunt similare, dar bitul de transport din poziția cea mai semnificativă este deplasat în poziția celui mai puțin semnificativ bit.

De exemplu, $\frac{4}{8}-\frac{3}{8}=\frac{1}{8}$. În formatul complement față de unu, transportul din MSB, dacă este prezent, este purtat spre LSB. Astfel, calculul $\frac{4}{8}-\frac{3}{8}=\frac{1}{8}$ devine $0{,}100 \oplus 1{,}100 = 0{,}000 \oplus 0{,}001 = 0{,}001$.

Adunarea şi scăderea în sistemul mărime cu semn sunt mai complexe şi, ca urmare, acesta este folosit mai mult la multiplicare, care se efectuează prin multiplicarea modulelor şi stabilind semnul produsului.

Exemplul 5.3.

Să se multiplice numerele 0,625 şi 0,250 folosind reprezentarea mărime cu semn.

Soluție.

Zecimal	Mărime cu semn
0,625	0,101
0,250	0,010
0000	000
3125	101
1250	000
0,156250	0,001010 = 0,156250

Multiplicarea în aritmetica complement fată de 1 şi față de 2 este mai dificilă şi necesită un hard sau algoritmi speciali.

Dacă rezultatul unei operații aritmetice depășește numărul maxim ce poate fi reprezentat pe *b* biți, apare *depășirea.* În procesarea digitală se folosește, de obicei, formatul fracționar, numerele care reprezintă mărimile ce intervin în procesare şi rezultatele operațiilor aritmetice sunt scalate, astfel încât modulul lor să nu depășească valoarea 1.

La multiplicarea numerelor fracționare, nu există probleme de depășire în cele trei aritmetici. Depășirea poate apărea numai când suma numerelor fracționare este mai mare decât 1. Dacă depășirea apare într-o etapă intermediară a adunării, în final nu va exista depășire, cu condiția ca valoarea absolută a rezultatului final să fie subunitară.

Exemplul 5.4.

Să se adune 0,3125 + 0,7500 + (-0,6250) folosind aritmetica în complement față de 1 pe cinci biți.

Soluție.

zecimal	complement față de 1	
0,3125	0,0101	
+0,7500	0,1100	
1,0625	1,0001	→ incorect, MSB = 1 implică număr negativ
-0,6250	1,0101	
0,4375	0,0111	→ ultimul 1 se datorează transportului

Exemplul 5.5.

Să se exprimă fracțiile $\frac{7}{8}$ şi $-\frac{7}{8}$ în formatele: mărime cu semn, complement față de 1 şi complement față de 2.

Soluție. $x = \frac{7}{8}$, este reprezentat ca $2^{-1}+2^{-2}+2^{-3}$, care, în formatul mărime cu semn conduce la $x = 0{,}111$, iar $x = -\frac{7}{8}$ este reprezentat ca x = 1,111. Reprezentarea în complement față de unu şi față de doi a lui $x = \frac{7}{8}$ este aceeaşi ca formatul mărime cu semn, adică $x = 0{,}111$. Reprezentarea în complement față de unu a lui $x = -\frac{7}{8}$ este x_{1C} = 1,000 şi în complement față de doi este $x_{2C} = 1{,}000 + 0{,}001 = 1{,}001$.

Deşi sunt posibile o mare varietate de alte reprezentări în virgulă fixă, cele descrise anterior sunt cele mai utilizate în practică. Cele mai multe procesoare de semnal în virgulă fixă folosesc aritmetica în complement faţă de doi. Aritmetica complementului faţă de doi este de fapt aritmetica modulo-2^{b+1} (adică orice număr care depăşeşte domeniul, este redus la acest domeniu, prin scăderea celui mai apropiat multiplu de 2^{b+1}).

La adunarea sau scăderea a două numere în virgulă fixă, fiecare de b biţi lungime (cu un bit adiţional de semn), rezultatul este un număr de b biţi. Dacă rezultatul adunării depăşeşte cel mai mare număr care poate fi reprezentat pe b biţi, apare depăşirea. Singura metodă pentru evitarea acestei probleme este creşterea numărului de biţi din acumulator şi, prin urmare, creşterea gamei dinamice care poate fi acoperită.

În general, înmulţirea a două numere în virgulă fixă, fiecare în lungime de b biţi, are ca rezultat un produs de lungime $2b$ biţi. În aritmetica cu virgulă fixă, produsul este de obicei trunchiat sau rotunjit la b biţi, ceea ce conduce la o eroare de trunchiere sau rotunjire cauzată de eliminarea celor mai puţin semnificativi b biţi.

Depăşirea în cazul adunării numerelor în reprezentarea în aritmetica în virgulă fixă este un dezavantaj cauzat de domeniul dinamic redus. Aritmetica în virgulă mobilă nu prezintă acest dezavantaj.

5.2.2. Reprezentarea numerelor în virgulă mobilă

Reprezentarea în virgulă fixă a numerelor, permite acoperirea unui domeniu dinamic, $x_{\max}$-$x_{\min}$ cu o rezoluţie

$$\Delta = \frac{x_{\max} - x_{\min}}{m-1}, \tag{5.16}$$

unde m=2^{b+1} este numărul de nivele, iar b+1 numărul de biţi. O caracteristică de bază a reprezentării în virgulă fixă este că rezoluţia este fixă. În plus, Δ creşte direct proporţional cu creşterea domeniului dinamic.

Reprezentarea în virgulă mobilă poate fi folosită ca o metodă de acoperire a unui domeniu dinamic mai larg. Reprezentarea în virgulă mobilă cel mai des întâlnită în practică constă dintr-o mantisă M, care este partea fracţionară a numărului şi se încadrează în domeniul $1/2 \le M < 1$, înmulţită cu factorul exponenţial 2^E unde exponentul E este un întreg pozitiv sau negativ. Un număr X, este reprezentat ca: $X = M \cdot 2^E$.

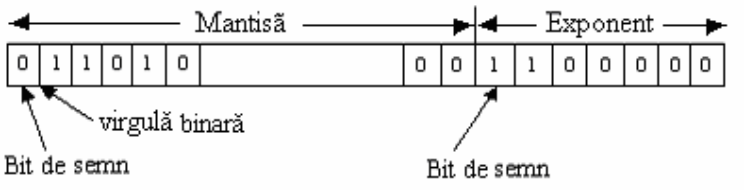

Figura 5.2 Reprezentarea în virgulă mobilă

Mantisa şi exponentul necesită fiecare câte un bit de semn pentru reprezentarea numerelor pozitive sau negative. Deoarece mantisa este o fracție cu semn, se poate folosi oricare din reprezentările în virgulă fixă descrise anterior.

De exemplu, numărul X_1=5 este reprezentat de următoarea mantisă si exponent:

M_1=0,101000

E_1=011

în timp ce numărul $X_2=\frac{3}{8}$ este reprezentat de următoarea mantisă şi exponent:

M_2=0,110000

E_2=101

Dacă cele două numere se înmulțesc, mantisele sunt înmulțite şi exponenții adunați. Prin urmare produsul celor două numere date mai sus este:

$$X_1 \cdot X_2 = M_1 \cdot M_2 \cdot 2^{E_1+E_2} = (0{,}011110) \cdot 2^{010} = (0{,}111100) \cdot 2^{001}$$

Împărțirea a două numere reprezentate în virgulă mobilă se efectuează prin împărțirea mantiselor şi scăderea exponenților.

$$\frac{X_1}{X_2} = \frac{M_1}{M_2} \cdot 2^{(E_2-E_2)}$$

Adunarea a două numere în virgulă mobilă necesită ca exponenții să fie egali. Aceasta se poate obține deplasând virgula binară a mantisei celui mai mic număr spre stânga şi compensând prin creşterea corespunzătoare a exponentului. Atunci numărul X_2 poate fi exprimat în forma

M_2=0,000011

E_2=011

Cu E_1=E_2, se pot aduna cele două numere X_1 şi X_2. Rezultatul este

$X_1 + X_2 = (0{,}101011) \cdot 2^{011}$

Se observă că operația de deplasare, impusă de egalarea exponenților lui X_2 şi X_1, poate conduce la o precizie mai mică în reprezentarea lui X_2. În exemplul anterior, mantisa pe şase biți a fost suficient de lungă pentru a se face deplasarea a patru biți la dreapta pentru M_2, fără a pierde nici unul. Totuşi o deplasare a cinci biți va cauza pierderea unui singur bit iar deplasarea a şase biți va conduce la mantisa M_2=0,000000; de aceea aceasta va trebui rotunjită după deplasare astfel încât M_2=0,000001.

Eroarea de depăşire apare la multiplicarea a două numere în virgulă mobilă când suma exponenților depăşeşte domeniul dinamic al reprezentării în virgulă fixă a exponentului.

Comparând reprezentarea în virgulă fixă cu cea în virgulă mobilă, cu acelaşi număr total de biți, rezultă că reprezentarea în virgulă mobilă permite acoperirea unui domeniu mai larg prin varierea rezoluției în acel interval. Rezoluția scade odată cu creşterea mărimii numerelor succesive. Cu alte cuvinte, distanța succesivă dintre două numere reprezentate în virgulă mobilă creşte odată cu creşterea numerelor în mărime. Astfel, pentru acoperirea aceluiaşi domeniu dinamic cu ambele reprezentări, în virgulă fixă şi virgulă mobilă, reprezentarea în virgulă mobilă oferă rezoluție fină pentru numere mici, dar rezoluție slabă pentru numere mari, spre deosebire de reprezentarea în virgulă fixă, care oferă o rezoluție uniformă în reprezentarea numerelor.

De exemplu, pentru un calculator care lucrează pe 32 biți, este posibilă reprezentarea a 2^{32} numere. Dacă se doreşte reprezentarea întregilor pozitivi începând cu zero, cel mai mare număr întreg ce poate fi reprezentat este: 2^{32}-1=4.294.967.295. Distanța dintre două numere succesive (rezoluția) este 1. Altfel, se poate folosi bitul cel mai din stânga ca bit de semn şi ceilalți 31 de biți rămaşi pentru valoare. Într-un astfel de caz reprezentarea în virgulă fixă permite acoperirea domeniului

$-(2^{31}-1)= -2.147.483.647$ la $(2^{31}-1)= 2.147.483.467$

tot cu o rezoluție de 1. Dacă, însă, se alocă 10 biți pentru partea fracționară, 21 de biți pentru partea întreagă şi un bit pentru semn, această reprezentare permite acoperirea domeniul dinamic:

$-(2^{31}-1) \cdot 2^{-10} = -(2^{21} - 2^{-10})$ la $(2^{31}-1) \cdot 2^{-10} = 2^{21} - 2^{-10}$ adică

de la -2.097.151,999 la 2.097.151,999

În acest caz, rezoluția este 2^{-10}. Prin urmare domeniul dinamic a fost scăzut cu un factor de aproximativ 1000 (2^{10} mai exact), în timp ce rezoluția a crescut cu același factor.

Pentru comparație, se presupune că cei 32 biți ai cuvântului sunt folosiți pentru a reprezenta numere în virgulă mobilă astfel: mantisa pe 23 de biți plus un bit de semn și exponentul cu 7 biți plus un bit de semn. Cel mai mic număr, în modul, va avea reprezentarea:

semn	23 biți	semn	7 biți	
0,	100.....0	1	1111111	$= \frac{1}{2} \times 2^{-127} \approx 0{,}3 \times 10^{-38}$

În cealaltă extremă, cel mai mare număr care poate fi reprezentat cu acest format în virgulă mobilă este:

semn	23 biți	semn	7 biți	
0,	11.....1	0	1111111	$= (1-2^{-23}) \times 2^{127} \approx 1{,}7 \times 10^{38}$

S-a obținut un domeniu dinamic de aproximativ 10^{76}, dar cu o rezoluție variabilă, adică rezoluție fină pentru numere mici și rezoluție slabă pentru numere mari.

5.2.3. Reprezentarea în virgulă mobilă pe bloc

Acest mod de reprezentare a numerelor este un hibrid între sistemele cu virgulă fixă și cele cu virgulă mobilă. În acest caz, în loc ca fiecare număr să fie reprezentat individual, ca în cazul sistemelor cu virgulă mobilă, un bloc sau un șir de numere are un exponent fix asociat. Acest exponent fix este obținut din examinarea tuturor numerelor din bloc și reprezentarea celui mai mare număr ca un număr cu virgulă mobilă cu o mantisă normalizată. Avantajul unui astfel de sistem constă în folosirea unui singur exponent pentru un bloc mare de numere. Astfel sistemul este potrivit pentru implementarea algoritmilor ce necesită un volum mare de calcule.

5.3. Efectele cuantizării în conversia A/D a semnalelor

Operațiile de bază îndeplinite de un convertor A/D sunt:

1. Să eșantioneze semnalul în mod periodic și cu rată de eșantionare suficient de mare pentru a evita eroarea alias;
2. Să cuantizeze amplitudinea eșantioanelor într-un set discret de nivele.

Prin urmare, dintr-un semnal analogic $x_a(t)$ eşantionat cu frecvenţa $F_s=1/T$, unde T este perioada de eşantionare, va rezulta o secvenţă $x[n]=x_a(nT)$, a cărei amplitudine este cuantizată, rezultând secvenţa

$$x_q[n] \equiv Q[x[n]] \tag{5.17}$$

unde $x_q[n]$ reprezintă semnalul cuantizat, iar $Q[\bullet]$ operaţia de cuantizare.

Dacă un semnal al cărui domeniu dinamic este R urmează a fi reprezentat pe $N=b+1$ biţi, numărul nivelelor de cuantizare ce pot fi reprezentate este de 2^{b+1}. În reprezentarea în virgulă fixă b biţi dau 2^b valori ale amplitudinii iar un bit dă informaţia de semn. Distanţa dintre două nivele adiacente sau pasul de cuantizare este $\Delta = \frac{R}{2^{b+1}}$ [63].

În reprezentarea în virgulă fixă a numerelor fracţionare, dacă domeniul dinamic depăşeşte ± 1, de multe ori este necesară scalarea semnalului, caz în care pasul de cuantizare al semnalului scalat este redus corespunzător la $\Delta_1 = \frac{2}{2^N} = 2^{-b}$.

Exemplul 5. 6.

Să se determine nivelele de cuantizare ale unui semnal continuu cu domeniul dinamic ± 20V după ce a fost eşantionat şi apoi procesat cu un convertor A/D pe N=4 biţi.

Soluţie. Pasul de cuantizare pentru semnalul nescalat este $\Delta = \frac{40}{2^4} = 2{,}5V$. Pasul de cuantizare pentru semnalul scalat la domeniul ± 1 este $\Delta_1 = \frac{2}{2^4} = 0{,}125V$ care este $2^{-b} = 2^{-3}$, adică valoarea corespunzătoare unui 1 în poziţia bitului cel mai puţin semnificativ.

5.3.1. Cuantizarea semnalului de intrare. Erori rezultate din rotunjire şi trunchiere

În executarea calculelor folosind aritmetica în virgulă fixă sau mobilă, apare problema cuantizării numerelor prin trunchiere sau rotunjire de la o reprezentare pe un anumit număr de biţi b_n (posibil a fi, la limită, şi infinit în cazul unui eşantion al unui semnal analogic) la o alta, pe un număr mai mic de biţi, b. Dacă valoarea semnalului se află între două nivele, aceasta poate fi aproximată fie prin cel mai apropiat nivel superior,

fie prin cel mai apropiat nivel inferior. Efectul cuantizării este că introduce o eroare a cărei valoare depinde de numărul de biți din numărul original şi de numărul de biți de după cuantizare.

Sunt trei metode de cuantizare frecvent folosite:

- *Rotunjirea,* caz în care valoarea semnalului este aproximată de cel mai apropiat nivel de cuantizare.
- *Trunchierea,* caz în care valoarea semnalului este aproximată de cel mai mare nivel care este inferior sau egal valoric cu eşantionul semnalului.
- *Trunchierea semn – valoare,* care este asemănătoare cu trunchierea pentru numere pozitive, dar valorile negative ale semnalului sunt aproximate de cel mai apropiat nivel de cuantizare mai mare sau egal cu semnalul.

Aceste descrieri se aplică cuantizării în aritmetica în virgulă fixă. Cele două metode de trunchiere rezultă din tratările diferite ale numerelor negative în reprezentările: mărime cu semn, complement fată de 1, complement fată de 2.

La un moment dat, nT, eroarea datorată cuantizării este

$$E_i = Q_i[x[n]] - x_a(nT) = x_{qi} - x_a \qquad (5.18)$$

unde $i = r$ în cazul rotunjirii şi $i = t$ în cazul trunchierii, $x_a = x_a(nT)$ reprezintă valoarea necuantizată a semnalului reprezentată pe $b_n + 1$ biți, iar $Q_i[x[n]] = x_{qi}$, valoarea cuantizată a semnalului reprezentată pe $b+1$ biți.

Rotunjirea

În cazul rotunjirii

$$E_r = Q_r[x[n]] - x_a(nT) = x_{qr} - x_a \text{ şi } \frac{-\Delta}{2} \le E_r \le \frac{\Delta}{2},\ \Delta = 2^{-b} \qquad (5.19)$$

Relația neliniară dintre x_{qr} şi x_a este reprezentată în figura 5.3 unde x_a este un semnal cu amplitudine continuă ($b_n = \infty$).

În *reprezentarea în virgulă fixă*, eroarea de rotunjire satisface relația (5.19), indiferent de aritmetica folosită pentru reprezentarea numerelor negative, deoarece rotunjirea este independentă de semn, ea depinzând numai de mărimea numărului.

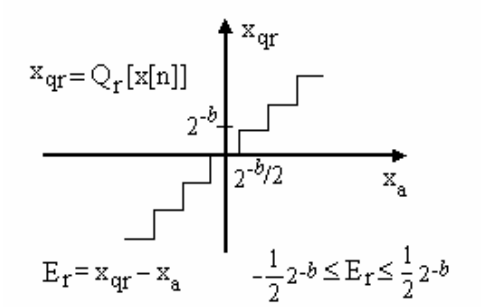

Figura 5.3 Relaţia dintre valorile cuantizate şi necuantizate în cazul rotunjirii

În *reprezentarea în virgulă mobilă*, mantisa este cea trunchiată sau rotunjită.

Dacă $$x_a = M_a \cdot 2^E \tag{5.20}$$

şi $$Q_r[x[n]] = M \cdot 2^E \tag{5.21}$$

atunci $$E_r = Q_r[x[n]] - x_a = (M - M_a)2^E \tag{5.22}$$

Dar pentru rotunjire $$-\Delta/2 \le M - M_a \le \Delta/2 \tag{5.23}$$

şi atunci din relaţia (5.19) rezultă

$$-2^E \Delta/2 \le E_r \le 2^E \Delta/2\,, \tag{5.24}$$

care dă eroarea absolută în virgulă mobilă datorată cuantizării mantisei.

Se defineşte *eroarea relativă* ε, astfel încât

$$Q_r[x[n]] = x_a(1+\varepsilon) \tag{5.25}$$

Datorită rezoluţiei neuniforme, eroarea corespunzătoare reprezentării în virgulă mobilă este proporţională cu numărul, adică

$$E_r = \varepsilon \cdot x_a \tag{5.26}$$

şi relaţia (5.24) devine

$$-2^E \Delta/2 \le \varepsilon\, x_a \le 2^E \Delta/2 \tag{5.27}$$

sau

$$-2^E \Delta/2 \le \varepsilon M_a 2^E \le 2^E \Delta/2 \tag{5.28}$$

adică

$$-\Delta/2 \le \varepsilon M_a \le \Delta/2 \tag{5.29}$$

Mantisa satisface relaţia

$$\frac{1}{2} \le M_a < 1 \tag{5.30}$$

Dacă $M_a = \frac{1}{2}$, din (5.29) se obține domeniul maxim al erorii relative ca fiind

$$-\Delta \le \varepsilon \le \Delta \tag{5.31}$$

Trunchierea

Dacă metoda de cuantizare este trunchierea, numărul este aproximat în aritmetica în virgulă fixă, prin cel mai mare nivel care este mai mic sau egal cu valoarea semnalului. Trunchierea numerelor pozitive, negative şi relația neliniară dintre x_{qt} şi x_a sunt reprezentate în figura 5.4, unde x_a este un semnal cu amplitudine continuă.

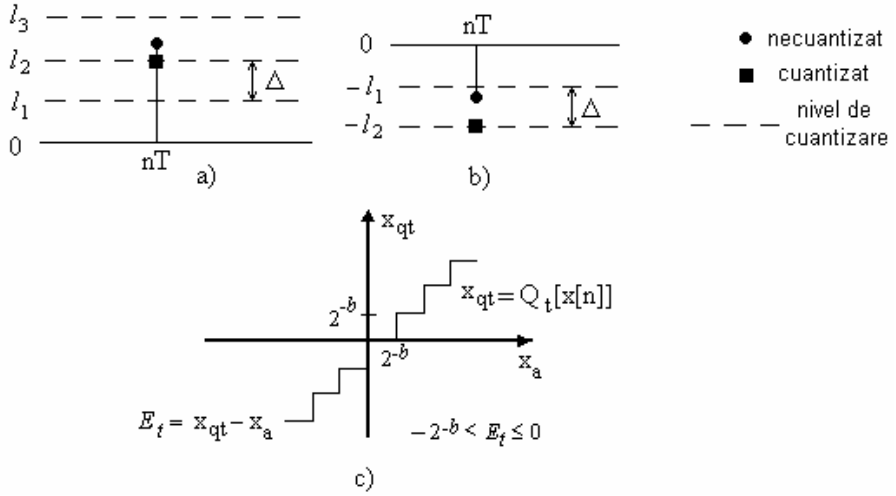

Figura 5.4. Relația dintre valorile cuantizate şi necuantizate în cazul trunchierii a) pentru numere pozitive, b) pentru numere negative, c) caracteristica de trunchiere în complement față de 2

Eroarea de trunchiere $E_t = Q_t[x[n]] - x_a$ este negativă sau zero.

$$-\Delta < E_t \le 0 \tag{5.32}$$

Acest lucru este valabil pentru toate numerele pozitive reprezentate în formatul mărime cu semn, complement față de 1 şi complement față de 2.

În continuare se examinează trunchierea numerelor negative reprezentate în diverse formate. Fie întâi reprezentarea în *complement față de* 2. Se consideră că numărul ce urmează a fi trunchiat este reprezentat

pe b_n+1 biți (la limită, se poate considera că $b_n = \infty$ pentru eşantioane ale unui semnal analogic). Modulul acestui număr negativ este

$$A_1 = 1 - \sum_{i=1}^{b_n} b_i \cdot 2^{-i} \tag{5.33}$$

Dacă acesta este trunchiat la *b* biți, modulul numărului devine

$$A = 1 - \sum_{i=1}^{b} b_i \cdot 2^{-i} \tag{5.34}$$

Diferența de mărime a modulului numărului negativ rezultată prin trunchiere este

$$A - A_1 = \sum_{i=1}^{b_n} b_i \cdot 2^{-i} - \sum_{i=1}^{b} b_i \cdot 2^{-i} = \sum_{i=b+1}^{b_n} b_i \cdot 2^{-i} \geq 0 \tag{5.35}$$

Deoarece modulul creşte prin trunchiere, numărul negativ reprezentat în complement față de 2 devine mai mic. Valoarea maximă a modulului erorii se obține când toți coeficienții b_i sunt egali cu 1, caz în care

$$A - A_1 = 2^{-b} - 2^{-b_n} < \Delta, \tag{5.36}$$

deoarece $\Delta = 2^{-b}$. Prin urmare, în reprezentarea în complement fată de 2, eroarea se situează în domeniul

$$-\Delta < E_t \leq 0 \tag{5.37}$$

Situația descrisă anterior este reprezentată în figura 5.4.

În cazul reprezentării numerelor negative în *complement față de* 1 pe b_n +1 biți, modulul numărului negativ este

$$A_1 = 1 - \sum_{i=1}^{b_n} b_i 2^{-i} - 2^{-b_n} \tag{5.38}$$

Prin trunchierea la *b*+1 biți, modulul numărului negativ devine

$$A = 1 - \sum_{i=1}^{b} b_i 2^{-i} - 2^{-b}, \tag{5.39}$$

astfel încât diferența acestora este

$$\begin{aligned} A - A_1 &= \sum_{i=1}^{b_n} b_i 2^{-i} - \sum_{i=1}^{b} b_i 2^{-i} + 2^{-b_n} - 2^{-b} = \\ &= \sum_{i=b+1}^{b_n} b_i 2^{-i} - \left(2^{-b} - 2^{-b_n}\right) \leq 0 \end{aligned} \tag{5.40}$$

Modulul numerelor negative descreşte prin trunchiere, adică, de fapt, acestea cresc. Situația este ilustrată în Figura 5.5. care reprezintă

trunchierea în reprezentarea semn - valoare. Prin urmare, domeniul în care poate lua valori eroarea ce apare prin trunchierea numerelor negative reprezentate în complement față de 1 este

$$0 \le E_t < \Delta \tag{5.41}$$

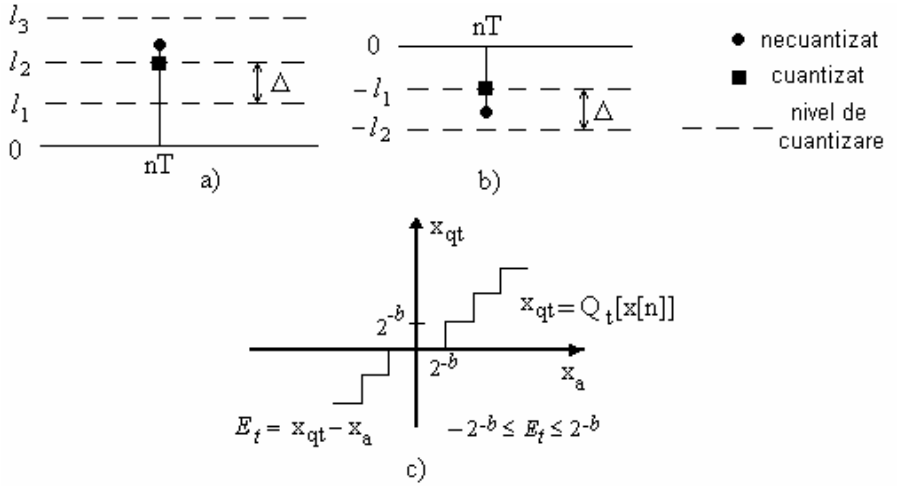

Figura 5.5. Relația dintre valorile cuantizate şi necuantizate în cazul trunchierii semn valoare a) numere pozitive, b) numere negative, c) caracteristica de trunchiere în semn – valoare

În reprezentarea numerelor negative în formatul *mărime cu semn*, biții care reprezintă modulul numărului negativ sunt aceeaşi cu cei corespunzători numărului pozitiv, diferind numai bitul de semn. Aceasta înseamnă că prin trunchierea unui număr negativ modulul acestuia scade, iar valoarea trunchiată este dată de cel mai apropiat nivel de cuantizare care nu este mai mic decât numărul, situație reprezentată în Figura 5.5.

În continuare se va considera trunchierea mantisei în cazul reprezentării în *virgulă mobilă*.

$$E_t = Q_t[x[n]] - x_a = (M - M_a)2^E \tag{5.42}$$

În reprezentarea în *complement față de* 2 a mantisei

$$-\Delta < M - M_a \le 0 \tag{5.43}$$

sau

$$-2^E \Delta < E_t \le 0 \tag{5.44}$$

Deoarece $E_t = \varepsilon x_a$, se obține

$$-2^E \Delta < \varepsilon x_a \le 0 \tag{5.45}$$

sau $$-2^E\Delta<\varepsilon M_a 2^E\le 0 \tag{5.46}$$

care implică $$-\Delta<\varepsilon M_a\le 0 \tag{5.47}$$

Dacă $M_a=\frac{1}{2}$ se obține domeniul maxim al erorii relative ε, ca fiind

$$-2\Delta<\varepsilon\le 0 \tag{5.48}$$

Dacă $M_a=-\frac{1}{2}$, domeniul erorii relative este

$$0\le\varepsilon<2\Delta \tag{5.49}$$

În reprezentarea în *complement față de* 1, eroarea de trunchiere pentru valori pozitive ale mantisei este:

$$-\Delta<M-M_a\le 0 \tag{5.50}$$

sau $$-2^E\Delta<E_t\le 0 \tag{5.51}$$

Cu $$E_t=\varepsilon x_a=\varepsilon M_a 2^E \tag{5.52}$$

şi $M_a=\frac{1}{2}$ se obține domeniul maxim al erorii relative pentru M_a pozitiv, ca fiind

$$-2\Delta<\varepsilon\le 0 \tag{5.53}$$

Pentru valori negative ale mantisei, eroarea este

$$0\le M-M_a<\Delta \tag{5.54}$$

sau $$0\le E_t<2^E\Delta \tag{5.55}$$

Pentru $M_a=-\frac{1}{2}$, domeniul maxim pentru eroarea relativă este

$$-2\Delta<\varepsilon\le 0, \tag{5.56}$$

aceeaşi ca şi pentru M_a pozitiv.

Acest lucru este valabil, de asemenea, şi pentru cazul în care mantisa este reprezentată în formatul *mărime cu semn*.

5.3.2. Model statistic pentru cuantizarea fină

În calculele aritmetice ce implică cuantizare prin trunchiere sau rotunjire, este convenabil să se adopte o metodă statistică pentru caracterizarea erorilor rezultate. Cuantizorul poate fi modelat prin

introducerea unui zgomot aditiv $e[n]$ ce se suprapune peste semnalul $x[n]$, cu respectarea unor ipoteze ce vor fi specificate în cele ce urmează, adică

$$Q[x[n]] = x_q[n] = x[n] + e[n] \tag{5.57}$$

unde $e[n]= E_r$ pentru rotunjire şi $e[n]= E_t$ pentru trunchiere, iar modelul este ilustrat în figura 5.6.

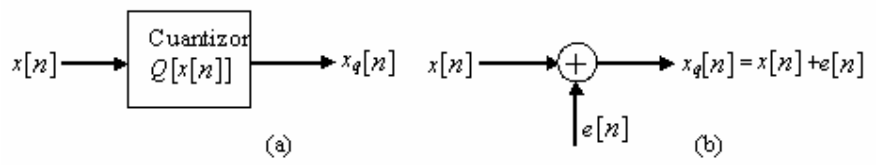

Figura 5.6. Modelul zgomotului aditiv pentru procesul liniar de cuantizare: (a) sistemul real; (b) model de cuantizare

Cum $x[n]$ poate fi orice număr care se încadrează în domeniul cuantizorului, eroarea de cuantizare este uzual modelată ca o variabilă aleatoare care se încadrează în limitele specificate anterior pentru erori. Mai mult, în practică, $b_n >> b$, deci mărimea 2^{-b_n} poate fi neglijată în relaţiile precedente. În aceste condiţii, erorile de cuantizare ale numerelor reprezentate în virgulă fixă şi virgulă mobilă se încadrează în intervalele prezentate în Tabelul 5.3.

Tabelul 5.3 Intervalele erorii de cuantizare

Tipul cuantizării	Tipul de aritmetică	Numere reprezentate cu virgulă fixă	Numere reprezentate cu virgulă mobilă
Rotunjire	-Semn-valoare -Complement faţă de 1 -Complement faţă de 2	$-2^{-b-1} \le E_r \le 2^{-b-1}$	$-2^{-b} \le \varepsilon \le 2^{-b}$
Trunchiere	Complement faţă de 2	$-2^{-b} < E_t \le 0$	$-2^{-b+1} < \varepsilon \le 0$, $x > 0$ $0 \le \varepsilon < 2^{-b+1}$, $x < 0$
Trunchiere semn-valoare	-Complement faţă de 1 -Semn-valoare	$-2^{-b} < E_t \le 0$, $x > 0$ $0 \le E_t < 2^{-b}$, $x < 0$	$-2^{-b+1} < \varepsilon \le 0$

În aceste condiții, funcțiile densitate de probabilitate pentru erorile de rotunjire şi trunchiere pentru formatele de reprezentare în virgulă fixă prezentate sunt ilustrate în figura 5.7 [49]. Se observă că în cazul trunchierii în formatul complement față de doi, valoarea medie a erorii are o deplasare de $2^{-b}/2$, în timp ce pentru celelalte cazuri ilustrate anterior, eroarea are o valoare medie nulă.

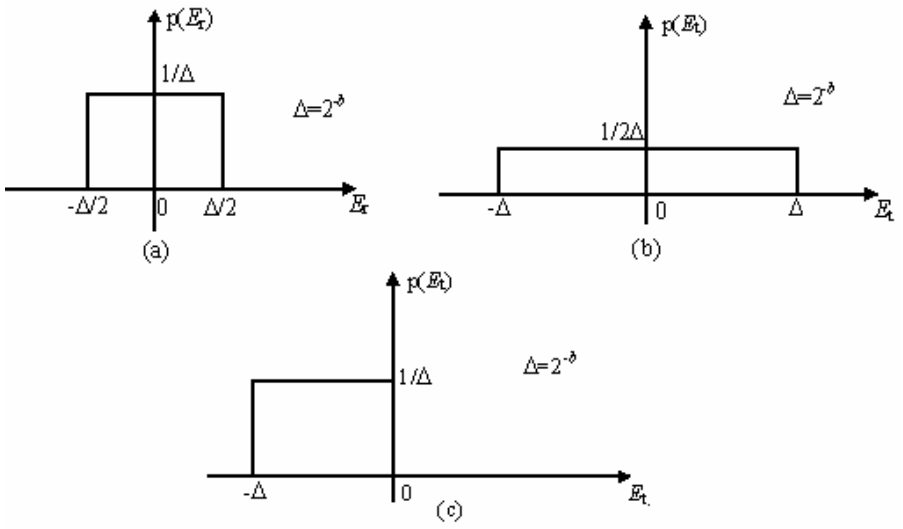

Figura 5.7 Caracterizarea statistică a erorilor de cuantizare. Funcțiile densitate de probabilitate ale (a) erorii de rotunjire; (b) erorii de trunchiere în formatul semn-valoare; (c) erorii de trunchiere în formatul complement față de doi

Analiza rezultatelor din Tabelul 5.3 şi a expresiilor densităților de repartiție pentru erorile de rotunjire şi trunchiere conduce la concluzia că rotunjirea este preferată altor metode de cuantizare, din următoarele motive[34]:

–semnalul de eroare este independent de tipul de aritmetică;

–media semnalului eroare este zero;

–nici o altă metodă de cuantizare nu conduce la o dispersie mai mică.

Cuantizarea reprezintă o operație neliniară şi ireversibilă.

Efectele erorii de cuantizare datorate rotunjirii pot fi evidențiate dacă $e[n]$ se consideră o secvență aleatoare care satisface următoarele proprietăți:

1. Eroarea $e[n]$ este uniform distribuită în domeniul $[-\Delta/2, \Delta/2]$,
2. Secvența de eroare $\{e[n]\}$ este o secvență de zgomot alb staționar, pentru care $e[n]$ şi $e[m]$, pentru $m \neq n$, sunt necorelate.
3. Secvența de eroare $\{e[n]\}$ este necorelată cu semnalul $x[n]$.

Ipotezele de mai sus sunt îndeplinite când pasul de cuantizare este mic şi semnalul $x[n]$ traversează mai multe nivele de cuantizare între două eşantioane succesive. Efectul zgomotului aditiv, $e[n]$, asupra semnalului dorit poate fi studiat evaluând raportul semnal-zgomot (SNR) care, pe scară logaritmică (în decibeli), este

$$SNR = 10 \cdot \log_{10} \frac{P_x}{P_n} \tag{5.58}$$

unde P_x este puterea semnalului, iar P_n este puterea zgomotului de cuantizare.

Dacă eroarea de cuantizare este uniform distribuită în domeniul (-Δ/2, Δ/2), aşa cum este reprezentat în figura 5.7a, valoarea medie a erorii este zero şi dispersia (puterea zgomotului de cuantizare) este

$$P_n = \sigma_e^{\,2} = \int_{-\Delta/2}^{\Delta/2} e^2 p(e) de = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} e^2 de = \frac{\Delta^2}{12} = \frac{2^{-2b}}{12} \tag{5.59}$$

Prin urmare, SNR este

$$SNR = 10 \cdot \log_{10} \frac{P_x}{P_n} = 10 \cdot \log_{10} P_x + 10 \cdot \log_{10}(12 \times 2^{2b}) \tag{5.60}$$

$$SNR = 10 \cdot \log_{10} P_x + 10{,}8 + 6b \tag{5.61}$$

Această expresie pentru SNR indică faptul că fiecare bit folosit în convertorul A/D sau cuantizor, măreşte raportul semnal/zgomot de cuantizare cu 6 dB sau reduce puterea zgomotului de cuantizare cu 6 dB.

De exemplu, dacă se stabileşte nivelul puterii zgomotului de cuantizare la –70 dB față de nivelul puterii semnalului, trebuie folosit un cuantizor pe 10 biți (sau convertor pe 10 biți).

Pentru a analiza efectul zgomotului de cuantizare asupra răspunsului unui sistem discret, liniar, invariant în timp, se consideră un astfel de sistem caracterizat de funcția pondere $h[n]$, la intrarea căruia se aplică semnalul cuantizat $x_q[n] = x[n] + e[n]$. Datorită liniarității sistemului, ieşirea sa este suma răspunsurilor sistemului la semnalul necuantizat $x[n]$ şi la eroarea de cuantizare $e[n]$. Notând semnalul de

ieşire datorat zgomotului sau erorii de cuantizare cu $z[n]$, conform figurii 5.8, se poate scrie

$$z[n] = \sum_{k=0}^{n} h[k]e[n-k] \tag{5.62}$$

relație din care poate fi determinată dispersia zgomotului de ieşire cauzat de eroarea de cuantizare.

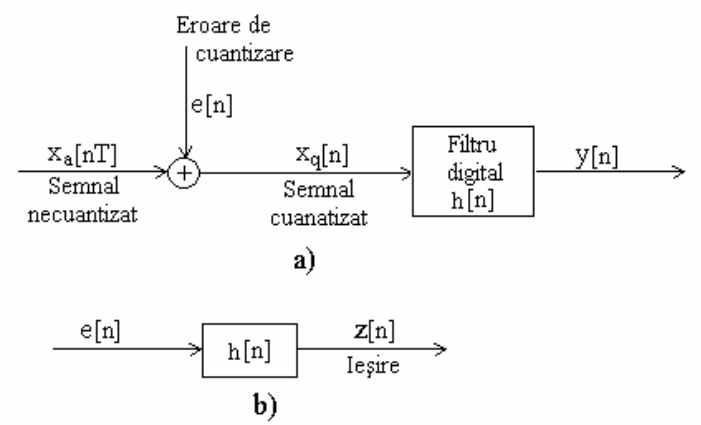

Figura 5.8. Model pentru eroarea datorată cuantizării semnalului de intrare
a) Modelul de eroare, b) ieşirea datorată zgomotului de cuantizare
b)

În cazul cuantizării prin rotunjire, ținând seama de ipotezele asumate pentru eroare şi de relația (5.59), dispersia fiecărui termen din suma (5.62) este

$$\sigma_e^2 h^2[k] = \frac{\Delta^2}{12} h^2[k] \tag{5.63}$$

Deoarece dispersia unei sume de variabile aleatoare independente este egală cu suma dispersiilor lor, rezultă că, în ipoteza că erorile de cuantizare s-au presupus independente la diferite momente de timp, dispersia ieşirii $z[n]$ este

$$\sigma_{0z}^2[n] = \frac{\Delta^2}{12} \sum_{k=0}^{n} h^2[k] \tag{5.64}$$

Dispersia creşte până la o valoare de regim permanent cu condiția ca filtrul să fie stabil. Dispersia de regim permanent se calculează cu relația

$$\sigma_{ozss}^2 = \lim_{n\to\infty} \sigma_{0z}^2[n] = \frac{\Delta^2}{12} \sum_{k=0}^{\infty} h^2[k] \tag{5.65}$$

O altă formă pentru expresia dispersiei de regim permanent a ieşirii poate fi obţinută cu ajutorul funcţiei de sistem a filtrului, $H(z)$, în felul următor:

$$H(z)=\sum_{k=0}^{\infty}h[k]z^{-k} \qquad (5.66)$$

$$H(z^{-1})=\sum_{m=0}^{\infty}h[m]z^{m} \qquad (5.67)$$

Prin urmare,

$$H(z)H(z^{-1})=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}h[k]h[m]z^{m-k} \qquad (5.68)$$

Multiplicând ambii membri cu z^{-1} şi integrând după z pe un contur închis ce conţine originea în planul z, rezultă

$$\int_c H(z)H(z^{-1})z^{-1}dz=\int_c\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}h[k]h[m]z^{m-k-1}dz \qquad (5.69)$$

Când conturul c este în regiunea de convergenţă pentru $H(z)$ şi $H(z^{-1})$, se poate schimba ordinea de sumare şi integrare din membrul drept. Se observă că cercul unitate este inclus în domeniul rezultat din intersecţia regiunilor de convergenţă pentru $H(z)$ şi $H(z^{-1})$, cu condiţia ca $H(z)$ să fie stabil. Astfel se justifică alegerea cercului unitate drept contur de integrare. Relaţia (5.69) devine

$$\int_c H(z)H(z^{-1})z^{-1}dz=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}h[k]h[m]\int_c z^{m-k-1}dz \qquad (5.70)$$

Deoarece conturul de integrare conţine originea planului Z, conform teoremei lui Cauchy [48]

$$\int_c z^{m-k-1}dz=\begin{cases}2\pi j & m=k\\ 0 & m\neq k\end{cases} \qquad (5.71)$$

Cu (5.71), relaţia (5.70) devine

$$\int_c H(z)H(z^{-1})z^{-1}dz=2\pi j\sum_{k=0}^{\infty}h^2[k] \qquad (5.72)$$

şi, deci,

$$\sum_{k=0}^{\infty}h^2[k]=\frac{1}{2\pi j}\int_c H(z)H(z^{-1})z^{-1}dz \qquad (5.73)$$

Din (5.65) şi (5.73) rezultă următoarea expresie pentru dispersia de regim permanent

$$\sigma_{ozss}^2 = \frac{\Delta^2}{12} \sum_{\substack{polii\ din \\ cercul\ unitate}} reziduurile\ lui\ H(z)H(z^{-1})z^{-1}, \qquad (5.74)$$

expresie care, de multe ori, este mai uşor de evaluat decât (5.65).

Exemplul 5. 7.

Să se determine dispersia de regim permanent a zgomotului de la ieşirea unui sistem cauzal, stabil, de ordinul întâi, datorat cuantizării semnalului de intrare.

Soluţie. Ecuaţia cu diferenţe care caracterizează sistemul este $y[n] = Ay[n-1] + x[n]$, cu $|A| < 1$. Răspunsul la impuls al acestui sistem este $h[n] = A^n u[n]$. Din (5.65) rezultă dispersia zgomotului de ieşire

$$\sigma_{oz}^2[n] = \frac{\Delta^2}{12} \sum_{k=0}^{n} A^{2k} = \frac{\Delta^2}{12} \frac{1 - A^{2(n+1)}}{1 - A^2}$$

Dispersia de regim permanent, când $n \to \infty$, este $\sigma_{0zss}^2 = \frac{\Delta^2}{12(1 - A^2)}$.

$H(z) = \frac{1}{1 - A z^{-1}}$, cu un pol în $z = A$, şi $H(z^{-1}) = \frac{1}{1 - A z}$ cu un pol în $z = \frac{1}{A}$ în afara cercului unitate. Conform relaţiei (5.74) rezultă

$$\sigma_{0zss}^2 = \frac{\Delta^2}{12} \sum \left(reziduurile\ lui\ \frac{z}{z - A} \cdot \frac{1}{1 - Az} \cdot z^{-1} \Big|_{z=A} \right) = \frac{\Delta^2}{12(1 - A^2)}$$

identică, evident, cu expresia obţinută anterior.

Pentru sisteme de ordin superior este mai uşor a se folosi relaţia (5.74) decât (5.65) din cauza complexităţii expresiei răspunsului la impuls.

5.4. Erori cauzate de cuantizarea coeficientilor filtrelor

5.4.1. Efectul cuantizării parametrilor filtrului asupra stabilităţii. Analiza senzitivităţii la cuantizarea coeficienţilor filtrelor IIR

Pentru a asigura stabilitatea unui filtru recursiv cauzal, toţi polii acestuia trebuie să fie în interiorul cercului unitate din planul *Z*. În multe

cazuri este de dorit ca un pol sau o pereche de poli să fie în apropierea cercului unitate. Dacă în acest caz pasul de cuantizare este atât de mare încât reprezentarea polilor să fie pe sau în afara cercului unitate, filtrul astfel implementat devine instabil.

Fie, de exemplu, un filtru de ordinul întâi

$$y[n] = A\,y[n-1] + x[n] \tag{5.75}$$

şi fie $N = b+1$, numărul biţilor disponibili reprezentării coeficientului A care, pentru un filtru stabil, este cuprins în domeniul $-1 < A < 1$.

Mărimea pasului de cuantizare este $\Delta = 2^{-b}$. Dacă $\varepsilon = 1 - A$ este distanţa de la pol la cercul unitate, cea mai mică valoare a lui ε care poate fi precis reprezentată este $\Delta = 2^{-b}$. Pentru asigurarea stabilităţii trebuie ca pasul de cuantizare să fie mai mic sau egal cu distanţa de la pol la cercul unitate, $\Delta \le \varepsilon$, adică $2^{-N+1} \le (1 - A)$, de unde rezultă

$$N \ge -\frac{\log_{10}\varepsilon}{\log_{10} 2} + 1 = -\frac{\log_{10}(1-A)}{\log_{10} 2} + 1 \tag{5.76}$$

Exemplul 5. 8.

a) Fie $A = e^{-aT}$, unde $a = 1 rad/s$, $T = 10^3$ secunde. Dacă se foloseşte trunchierea ca metodă de cuantizare, să se determine numărul minim de biţi, N, necesar reprezentării lui A, astfel încât să nu rezulte instabilitate.

b) Dacă sunt disponibili 9 biţi şi $T = 10^3$ secunde, să se găsească a, astfel încât filtrul să fie stabil.

Soluţie. a) $1 - A = 1 - e^{-aT} \approx aT$, prin urmare, $N \ge -\dfrac{\log_{10} aT}{\log_{10} 2} + 1 = 11$ biţi

b) $9 = -\dfrac{\log_{10}\left(10^{-3} \cdot a\right)}{0.3} + 1$ care necesită $a = 4$ rad/secundă .

Pentru filtrele de ordin superior localizarea polilor depinde, în general, de mai mulţi coeficienţi. Pentru a ilustra efectul cuantizării coeficienţilor asupra localizării polilor şi, implicit, asupra caracteristicii de frecvenţă, fie un filtru IIR cu funcţia de sistem

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}} \tag{5.77}$$

Filtrul IIR cu coeficienți cuantizați are funcția de sistem

$$\overline{H}(z) = \frac{\sum_{k=0}^{M} \overline{b}_k z^{-k}}{1 + \sum_{k=1}^{N} \overline{a}_k z^{-k}} \tag{5.78}$$

unde coeficienții cuantizați $\{\overline{b}_k\}$ și $\{\overline{a}_k\}$ pot fi exprimați în funcție de coeficienții necuantizați $\{b_k\}$ și $\{a_k\}$ prin relațiile

$$\begin{aligned} \overline{a}_k &= a_k + \Delta a_k \qquad k = 1,2,...,N \\ \overline{b}_k &= b_k + \Delta b_k \qquad k = 0,1,...M \end{aligned} \tag{5.79}$$

$\{\Delta b_k\}$ și $\{\Delta a_k\}$ reprezentând erorile de cuantizare ale coeficienților.

Numitorul lui $H(z)$ poate fi exprimat în forma

$$D(z) = 1 + \sum_{k=0}^{N} a_k z^{-k} = \prod_{k=1}^{N} (1 - p_k z^{-1}) \tag{5.80}$$

unde $\{p_k\}$ sunt polii lui $H(z)$. Similar, se poate descompune numitorul lui $\overline{H}(z)$ în forma

$$\overline{D}(z) = \prod_{k=1}^{N} (1 - \overline{p}_k z^{-1}) \tag{5.81}$$

unde $\overline{p}_k = p_k + \Delta p_k$, k=1, 2, ..., N, și Δp_k este eroarea sau perturbația care rezultă din cuantizarea coeficienților filtrului.

În continuare, se urmărește a se exprima perturbația totală Δp_i a polului p_i, în funcție de eroarea de cuantizare $\{\Delta a_k\}$ a coeficienților. Perturbația Δp_i poate fi exprimată ca [48]

$$\Delta p_i = \sum_{k=1}^{N} \frac{\partial p_i}{\partial a_k} \Delta a_k \tag{5.82}$$

unde $\frac{\partial p_i}{\partial a_k}$ reprezintă variația poziției polului p_i determinată de variația coeficientului a_k. Astfel, eroarea totală este exprimată ca o sumă a erorilor datorate schimbărilor în fiecare din coeficienții $\{a_k\}$.

Derivatele parțiale $\partial p_i / \partial a_k$, k=1, 2, ..., N, pot fi obținute diferențiind $D(z)$ în funcție de fiecare $\{a_k\}$, după cum urmează [48]:

$$\left(\frac{\partial D(z)}{\partial a_k} \right)_{z=p_i} = \left(\frac{\partial D(z)}{\partial z} \right)_{z=p_i} \left(\frac{\partial p_i}{\partial a_k} \right) \tag{5.83}$$

Din (5.83) rezultă

$$\frac{\partial p_i}{\partial a_k} = \frac{\left(\partial D(z) / \partial a_k\right)_{z=p_i}}{\left(\partial D(z) / \partial z\right)_{z=p_i}} \tag{5.84}$$

Numărătorul relației (5.84) este

$$\left(\frac{\partial D(z)}{\partial a_k}\right)_{z=p_i} = z^{-k}\Big|_{z=p_i} = p_i^{-k} \tag{5.85}$$

Numitorul relației (5.84) este

$$\left(\frac{\partial D(z)}{\partial z}\right)_{z=p_i} = \left\{\frac{\partial}{\partial z}\left[\prod_{l=1}^{N}(1-p_l z^{-1})\right]\right\}_{z=p_i} =$$

$$= \left\{\sum_{k=1}^{N}\frac{p_k}{z^2}\prod_{\substack{l=1\\ l\neq k}}^{N}(1-p_l z^{-1})\right\}_{z=p_i} = \frac{1}{p_i^{N}}\prod_{\substack{l=1\\ l\neq i}}^{N}(p_i - p_l) \tag{5.86}$$

Prin urmare, relația (5.84) poate fi exprimată sub forma

$$\frac{\partial p_i}{\partial a_k} = \frac{p_i^{N-k}}{\prod_{\substack{l=1\\ l\neq i}}^{N}(p_i - p_l)} \tag{5.87}$$

Înlocuind rezultatul din (5.87) în (5.82) rezultă eroarea totală de perturbație Δp_i în forma

$$\Delta p_i = \sum_{k=1}^{N}\frac{p_i^{N-k}}{\prod_{\substack{l=1\\ l\neq i}}^{N}(p_i - p_l)}\Delta a_k \tag{5.88}$$

Această expresie oferă o măsură a senzitivității polului p_i la o schimbare a coeficienților $\{a_k\}$.

Un rezultat analog se poate obține pentru senzitivitatea zerourilor la erorile cauzate de cuantizarea parametrilor $\{b_k\}$.

Termenii $(p_i - p_l)$ din numitorul relației (5.88) reprezintă vectori, în planul *Z,* orientați de la polii $\{p_l\}$ la polul $\{p_i\}$. Dacă polii sunt foarte grupați, ca în cazul unui filtru de bandă îngustă reprezentat în figura 5.9, lungimile $|p_i - p_l|$ vor fi mici pentru polii din vecinătatea lui p_i. Aceste lungimi mici vor contribui la erori mari şi va rezulta o perturbație Δp_i mare. Eroarea Δp_i poate fi minimizată prin maximizarea lungimii $|p_i - p_l|$.

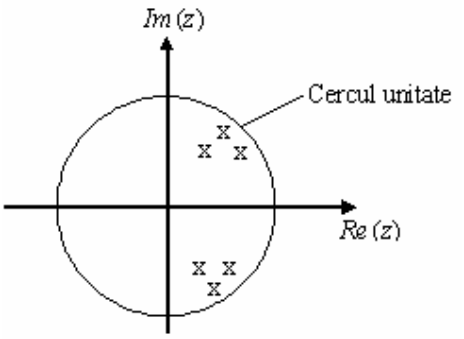

Figura 5.9 Poziții ale polilor unui filtru IIR trece bandă

Acest lucru se poate realiza prin implementarea filtrelor de ordin mare cu celule cu un singur pol sau cu doi poli. Fltrele cu un singur pol (şi un singur zero) au valori complexe pentru coeficienți şi necesită operații aritmetice în complex pentru realizarea lor. Această problemă poate fi evitată combinând polii şi zerourile complex conjugate, pentru a forma secțiuni de filtru de ordin doi cu coeficienți reali. Deoarece polii complex conjugați sunt suficient de depărtați, eroarea de cuantizare Δp_i este minimizată şi, în consecință, filtrul cu coeficienții cuantizați rezultat aproximează mai bine caracteristica răspunsului în frecvență a filtrului cu coeficienții necuantizați.

Exemplul 5. 9.

Un filtru digital de ordinul doi are polii reali p_1 şi p_2. Acesta este implementat în forma directă. Se cere:

a) Din relația generală (5.82) să se scrie o relație pentru modificarea poziției polilor datorată modificărilor coeficienților ecuației cu diferențe corespunzătoare.
b) Dacă $p_1 = 0{,}98$ şi $p_2 = 0{,}94$, care este numărul minim de biți necesar ca filtrul să rămână stabil în urma cuantizării coeficienților? Metoda de cuantizare se presupune a fi rotunjirea.

*Soluție***.** a) Din (5.87) rezultă $\dfrac{\partial p_i}{\partial a_k} = \dfrac{p_i^{2-k}}{\prod\limits_{\substack{l=1 \\ l \neq i}}^{2} (p_i - p_l)}$, $k = 1, 2$ şi $i = 1, 2$.

$$\frac{\partial p_1}{\partial a_1} = \frac{p_1}{p_1 - p_2} \qquad \frac{\partial p_1}{\partial a_2} = \frac{1}{p_1 - p_2}$$

$$\frac{\partial p_2}{\partial a_1} = \frac{p_2}{p_2 - p_1} \qquad \frac{\partial p_2}{\partial a_2} = \frac{1}{p_2 - p_1}$$

variația totală în poziția polilor este

$$\Delta p_i = \sum_{k=1}^{2} \frac{\partial p_i}{\partial a_k} \Delta a_k$$

adică $\Delta p_1 = \frac{\partial p_1}{\partial a_1}\Delta a_1 + \frac{\partial p_1}{\partial a_2}\Delta a_2 = \frac{1}{p_1 - p_2}\left[p_1 \Delta a_1 + \Delta a_2\right]$

şi $\Delta p_2 = \frac{\partial p_2}{\partial a_1}\Delta a_1 + \frac{\partial p_2}{\partial a_2}\Delta a_2 = \frac{1}{p_2 - p_1}\left[p_2 \Delta a_1 + \Delta a_2\right]$

b)Este necesar a determina Δa_1 şi Δa_2. Numitorul funcției de transfer a filtrului are forma $(z - p_1)(z - p_2) = z^2 - a_1 z + a_2$ unde $a_1 = p_1 + p_2$ şi $a_2 = p_1 p_2$. Pentru asigurarea stabilității trebuie ca $-2 < a_1 < 2$ şi $-1 < a_2 < 1$ [63]. În aritmetica în virgulă fixă coeficientul a_1 poate fi scalat pentru a se obține un număr fracționar, deşi pentru coeficienții filtrului virgula binară este adesea mutată spre dreapta pentru a adapta coeficienții la mărimi mai mari ca unitatea. În orice caz se poate calcula pasul de cuantizare şi numărul de biți, N=b+1.

Pentru a_1, $\Delta = \frac{4}{2^N}$ şi pentru rotunjire $\Delta a_1 = \frac{\Delta}{2} = \frac{2}{2^N}$

S-ar putea alege acelaşi pas de cuantizare şi pentru a_2, caz în care ar fi necesari N - 1 biți deoarece domeniul lui a_2 este jumătate din cel pentru a_1. În schimb, s-ar putea adopta N biți pentru ambele registre, pentru a_1 şi a_2 şi pasul de cuantizare pentru a_2 să fie $\frac{2}{2^N} = \frac{\Delta}{2}$, astfel încât, pentru rotunjire $\Delta a_2 = \Delta/4 = \frac{1}{2^N}$.

Pentru ultima alegere, din expresia menționată anterior pentru schimbarea poziției polului rezultă

$$\Delta p_1 = \frac{1}{0{,}98 - 0{,}94}\left[(0{,}98)2 + 1{,}0\right]/2^N = 74/2^N \quad \text{şi}$$

$$\Delta p_2 = \frac{1}{0{,}94 - 0{,}98}\left[(0{,}94)2 + 1{,}0\right]/2^N = -72/2^N$$

Polul p_1, fiind mai apropiat de cercul unitate este posibil să cauzeze instabilitatea filtrului, dacă nu este reprezentat adecvat. Pentru stabilitate, trebuie să fie îndeplinită relația $1-p_1 = 0{,}02 > \Delta p_1 = 74/2^N$ sau $2^N > 3700$, care implică N=12 biți lungimea minimă a registrului.

Pentru a completa analiza, este necesar a considera şi cazul polilor complex conjugați în expresia funcției de transfer (5.77). Numitorul acesteia se poate scrie

$$1+\sum_{k=1}^{N} a_k z^{-k} = \prod_{i=1}^{q}\left(1-p_i z^{-1}\right)\prod_{k=1}^{s}\left[1-2r_k\left(\cos\theta_k\right)z^{-1}+r_k^2 z^{-2}\right] \qquad (5.89)$$

unde $s=\dfrac{N-q}{2}$, cu q poli simpli şi s perechi de poli complex conjugați.

Diferențiind (5.89) în raport cu a_l, cu $1 \le l \le N$ se determină senzitivitatea la cuantizarea coeficienților $\dfrac{\partial p_m}{\partial a_l}$, $1 \le m \le q$ şi $\dfrac{\partial r_g}{\partial a_l}$, şi $\dfrac{\partial \theta_g}{\partial a_l}$, $1 \le g \le s$. După câteva prelucrări matematice rezultă pentru polii simpli p_m, $1 \le m \le q$ [58]

$$\frac{\partial p_m}{\partial a_l} = \frac{p_m^{-l+1}}{\prod\limits_{\substack{i=1\\ i\neq m}}^{q}\left(1-p_i p_m^{-1}\right)\prod\limits_{k=1}^{s}\left[1-2r_k\left(\cos\theta_k\right)p_m^{-1}+r_k^2 p_m^{-2}\right]} \quad , \qquad (5.90)$$

şi pentru polii complecşi $r_g e^{\pm\theta_g}$, $1 \le g \le s$

$$\frac{\partial r_g}{\partial a_l} = \frac{-r_g^{-l+1}\sin\left[(l-1)\theta_g\right]}{2C_g \sin\theta_g} \qquad (5.91)$$

$$\frac{\partial \theta_g}{\partial a_l} = \frac{r_g^{-l}\left\{\sin\left[(l-2)\theta_g\right]-\cos\theta_g \sin\left[(l-1)\theta_g\right]\right\}}{2C_g \sin^2\theta_g} \qquad (5.92)$$

unde

$$C_g = \prod_{i=1}^{q}\left(1-p_i z^{-1}\right)\prod_{\substack{k=1\\ k\neq g}}^{N}\left(1-2r_k\cos\theta_k z^{-1}+r_k^2 z^{-2}\right)\Big|_{z=r_g e^{j\theta_g}} \qquad (5.93)$$

Deviațiile totale sunt

$$\Delta p_i = \sum_{l=1}^{N} \frac{\partial p_m}{\partial a_l} \Delta a_l \qquad l = 1, \ldots, q \tag{5.94}$$

$$\Delta r_g = \sum_{l=1}^{N} \frac{\partial r_g}{\partial a_l} \Delta a_l \qquad g = 1, \ldots, s \tag{5.95}$$

$$\Delta \theta_g = \sum_{l=1}^{N} \frac{\partial \theta_g}{\partial a_l} \Delta a_l \qquad g = 1, \ldots, s \tag{5.96}$$

Din nou se observă că, dacă polii sunt grupați, ca în cazul filtrelor de bandă îngustă, polii realizării în forma directă sunt sensibili la erorile de cuantizare a coeficienților şi, cu cât este mai mare numărul de poli grupați, cu atât şi senzitivitatea este mai mare.

Este interesant de observat modul în care influențează structura de implementare a filtrului erorile cauzate de cuantizarea coeficienților. Pentru a ilustra acest lucru, fie un filtru cu doi poli complex conjugați, caracterizat de funcția de sistem

$$H(z) = \frac{1}{1 - (2r\cos\theta)z^{-1} + r^2 z^{-2}} \tag{5.97}$$

Filtrul are polii la $z_{1,2} = re^{\pm j\theta}$. Când este realizat ca în figura 5.10, există doi coeficienți: $a_1 = -2r\cos\theta$ şi $a_2 = r^2$. Cu precizie infinită este posibil să obținem un număr infinit de poziții ale polilor. Evident, cu precizie finită (adică a_1 şi a_2 cuantizați), pozițiile posibile ale polilor sunt în număr finit.

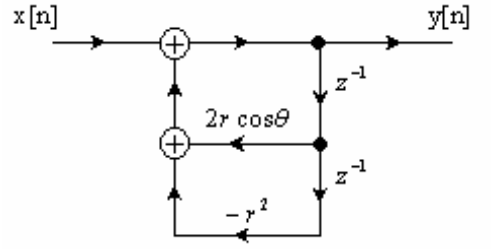

Figura 5.10. Realizare directă a unui filtru cu doi poli

De exemplu, pentru b=3, sunt posibile 7 valori nenule pentru a_1 şi a_2. În figura 5.11 sunt reprezentate pozițiile posibile ale polilor, numai pentru primul cadran al planului z. Sunt posibile 40 de poziții ale polilor în acest caz. Neuniformitatea în poziția polilor este datorată faptului că se cuantizeaza r^2 iar polii se găsesc pe un arc de cerc de rază r. Pentru o anumită cuantizare a coeficienților, polii se află pe o grilă din planul z

definită de intersecția cercurilor concentrice corespunzătoare cuantizării lui r^2 şi liniilor verticale corespunzătoare cuantizării lui $2r\cos\theta$. De importanță particulară este setul rar de poli, pentru θ apropiat de zero şi, datorită simetriei, pentru θ în apropierea lui π. Această situație va fi critic nefavorabilă pentru filtrele trece jos şi filtrele trece sus care au în mod normal polii grupați in jurul frecventei unghiulare θ=0 şi, respectiv, θ=π.

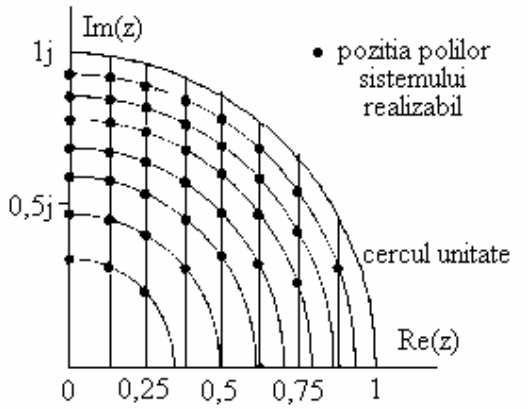

Fig. 5.11 Poziții posibile ale polilor structurii de ordinul doi în planul *Z*, pentru cuantizarea pe trei biți

O alternativă în realizarea filtrelor cu doi poli este forma cuplată, reprezentată în figura 5.12.

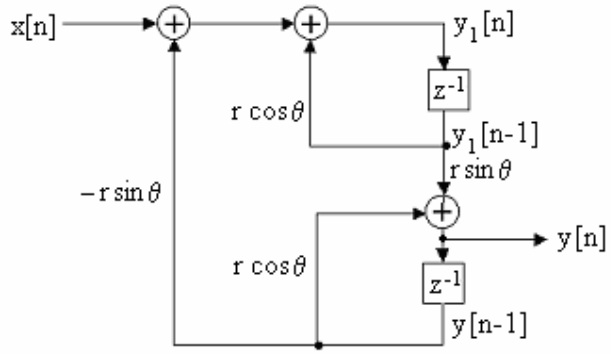

Figura 5.12. Realizare în forma cuplată a filtrului IIR cu doi poli

Cele două ecuații cuplate sunt:

$$\begin{aligned} y_1[n] &= x[n] + (r\cos\theta)\cdot y_1[n-1] - (r\sin\theta)\cdot y[n-1] \\ y[n] &= (r\sin\theta)\cdot y_1[n-1] + (r\cos\theta)\cdot y[n-1] \end{aligned} \qquad (5.98)$$

Transformând aceste ecuaţii în domeniul Z, se poate scrie

$$\frac{Y(z)}{X(z)} = H(z) = \frac{(r\sin\theta)z^{-1}}{1-(2r\cos\theta)z^{-1}+r^2z^{-2}} \tag{5.99}$$

În forma cuplată se observă că sunt de asemenea doi coeficienţi, $\alpha_1 = r\sin\theta$ şi $\alpha_2 = r\cos\theta$. Deoarece ambii sunt liniari în r, poziţiile posibile ale polilor sunt acum puncte egal spaţiate pe un caroiaj dreptunghiular, ca în figura 5.13.

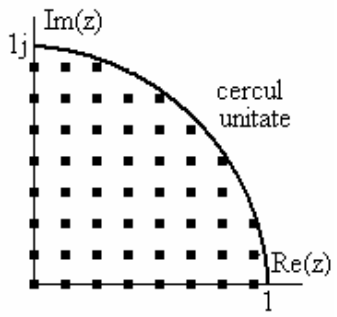

Figura 5.13. Poziţii posibile ale polilor filtrului cu doi poli, realizat în forma cuplată din figura 5.12

Ca urmare, poziţionarea polilor este acum uniform distribuită în interiorul cercului, lucru mult mai favorabil decât realizarea precedentă, mai ales pentru filtrele trece jos. Preţul plătit pentru această distribuire uniformă a poziţiei polilor este o creştere a volumului de calcule. Realizarea în formă cuplată necesită patru multiplicări, câte două pentru fiecare ieşire, în timp ce realizarea din figura 5.10 necesită doar două multiplicări. Este interesant de observat faptul că pentru o anumită lungime a coeficienţilor, forma directă permite o plasare mai adecvată a polilor cu r apropiat de unitate şi θ mare, pe când forma cuplată este mai avantajoasă pentru θ mic.

Deoarece sunt diverse metode de a realiza secţiunile de ordin doi ale filtrelor, este, de asemenea, clar că sunt multe posibilităţi pentru localizarea polilor în cazul coeficienţilor cuantizaţi. Ideal ar fi să se selecteze o structură care conduce la un set dens de puncte în regiunea unde se află polii. Din nefericire nu există o metodă simplă şi sistematică pentru determinarea realizării filtrului care să ducă la rezultatul dorit.

Având dat un filtru IIR de ordin înalt care trebuie implementat ca o combinaţie de secţiuni de ordinul doi, va trebui să se decidă între o structură în cascadă şi una în paralel, adică între realizarea

$$H(z) = \prod_{k=1}^{K} \frac{b_{k0} + b_{k1} z^{-1} + b_{k2} z^{-2}}{1 + a_{k1} z^{-1} + a_{k2} z^{-2}} \tag{5.100}$$

şi realizarea

$$H(z) = \sum_{k=1}^{K} \frac{c_{k0} + c_{k1} z^{-1}}{1 + a_{k1} z^{-1} + a_{k2} z^{-2}} \tag{5.101}$$

Dacă filtrul IIR are zerouri pe cercul unitate, cum este cazul filtrelor eliptice şi Cebyshev de tipul doi, fiecare secţiune de ordin doi din configuraţia în cascadă din (5.100) conţine o pereche de zerouri complex conjugate. Coeficienţii $\{b_{ki}\}$ din (5.100) determină în mod direct poziţiile acestor zerouri, iar cuantizarea lor tinde să le deplaseze de pe cercul unitate. Senzitivitatea răspunsului sistemului la eroarea de cuantizare este uşor şi direct controlabilă prin alocarea unui număr suficient de biţi pentru reprezentarea coeficienţilor cuantizaţi $\{b_{ki}\}$ cu o precizie specificată. Astfel va exista control direct asupra polilor şi zerourilor care rezultă din procesul de cuantizare. De fapt, se poate evalua efectul perturbării rezultate din cuantizarea coeficienţilor $\{b_{ki}\}$, cu o anumită precizie cerută.

Realizarea în paralel a lui $H(z)$, conform relaţiei (5.101), asigură un control direct doar asupra polilor sistemului. Coeficienţii numărătorului $\{c_{k0}\}$ şi $\{c_{k1}\}$ sunt obţinuţi prin descompunerea în fracţii simple a lui $H(z)$. Prin urmare polii influenţează indirect localizarea zerourilor, prin combinarea tuturor termenilor din descompunerea în fracţii simple a lui $H(z)$ şi, în consecinţă, este mult mai dificil a se determina efectul erorii de cuantizare datorat coeficienţilor $\{c_{ki}\}$, în localizarea zerourilor sistemelor.

Cuantizarea parametrilor $\{c_{ki}\}$ poate produce o perturbaţie semnificativă a poziţiilor zerourilor şi, de obicei, va fi suficient de mare în implementările cu virgulă fixă pentru a deplasa zerourile de pe cercul unitate. Aceasta este o situaţie foarte neplăcută, care poate fi însă remediată folosind o reprezentare în virgulă mobilă. În orice caz, structura în cascadă este mult mai robustă în prezenţa cuantizării coeficienţilor şi trebuie să fie alegerea preferată în aplicaţii practice, mai ales unde este folosită reprezentarea în virgulă fixă.

5.4.2. Cuantizarea coeficienților filtrelor FIR

Aşa cum s-a arătat şi în secțiunea precedentă, analiza senzitivității aplicată polilor unui sistem se aplică direct şi zerourilor filtrelor IIR. Prin urmare, o expresie asemănătoare cu relația (5.88) se poate obține pentru zerourile unui filtru FIR. Pentru a minimiza senzitivitatea la cuantizarea coeficienților, va trebui ca filtrul FIR cu un număr mare de zerouri să fie implementat ca o cascadă de secțiuni de ordinul unu şi doi.

Un aspect important în practică îl reprezintă filtrele FIR cu răspuns liniar de fază. Realizările directe ale unor astfel de filtre mențin proprietatea de fază liniară chiar şi în cazul cuantizării coeficienților. Aceasta rezultă din observația că funcția de sistem a unui filtru FIR de fază liniară satisface proprietatea

$$H(z) = \pm z^{-(M-1)} H(z^{-1}), \tag{5.102}$$

indiferent dacă coeficienții sunt sau nu, cuntizați.

Prin urmare, cuantizarea coeficienților filtrului FIR afectează doar caracteristica de amplitudine.

Din practică se ştie că pentru a reprezenta coeficienții unui filtru FIR de fază liniară de lungime moderată (M=32 ÷ 256) sunt necesari cel puțin 10 biți, dar, dacă este posibil, se preferă a se folosi 12 până la 14 biți. Cu creşterea lungimii filtrului trebuie să crească şi numărul de biți pentru reprezentarea coeficienților, pentru a menține aceeaşi eroare în răspunsul în frecvență al filtrului. Se presupune, de exemplu, că fiecare coeficient al filtrului este rotunjit la (b+1) biți. Prin urmare, eroarea de rotunjire se încadrează în domeniul: $-2^{-b}/2 < e_r[n] < 2^{-b}/2$.

Valoarea cuantizată a răspunsului la impuls poate fi reprezentată ca $h_q[n] = h[n] + e_r[n]$ şi eroarea în răspunsul în frecvență este

$$E_M(\omega) = \sum_{n=0}^{M-1} e_r[n] \cdot e^{-j\omega n} \tag{5.103}$$

Presupunând că $e_r[n]$ este o variabilă aleatoare uniform distribuită în intervalul $[-2^{-b}/2,\ 2^{-b}/2]$ cu valoarea medie zero, $E_M(\omega)$ va fi, de asemenea, de medie zero. Presupunând, în continuare, că $e_r[n]$ poate fi modelată ca o secvență de zgomot alb staționar, secvența erorilor $e_r[n]$, $0 \le n \le M-1$, are eşantioanele necorelate. Prin urmare, dispersia erorii în răspunsul în frecvență $E_M(\omega)$ este suma dispersiilor celor M termeni $e_r[n]$

$$\sigma_E^2 = \frac{2^{-2b}}{12} M \tag{5.104}$$

Ecuația (5.104) subliniază faptul că dispersia erorii creşte liniar cu lungimea filtrului *M.* Deviația standard a erorii $E_M(\omega)$ este

$$\sigma_E = \frac{2^{-b}}{\sqrt{12}}\sqrt{M} \tag{5.105}$$

Prin urmare, pentru fiecare creştere de patru ori a lui *M*, precizia în reprezentarea coeficienților filtrului trebuie crescută cu un bit, pentru a menține deviația standard fixă. Din practică se constată că pentru a avea o deviație standard acceptabilă se folosesc 12, 13 biți. Dacă lungimea filtrului, *M,* este mai mare decât 256 sau numărul de biți folosiți pentru reprezentarea coeficienților este mai mic de 12, atunci filtrul trebuie implementat ca o cascadă de secțiuni de filtre de lungimi mai mici.

Într-o realizare în cascadă, de forma

$$H(z) = G \cdot \prod_{k=1}^{K} H_k(z) \tag{5.106}$$

secțiunile de ordinul doi sunt:

$$H_k(z) = 1 + b_{k1}z^{-1} + b_{k2}z^{-2}. \tag{5.107}$$

Coeficienții au forma $b_{k1} = -2r_k \cos\theta_k$ şi $b_{k2} = r_k^{\,2}$. Cuantizarea lui b_{k1} şi b_{k2} conduce la localizarea zerourilor ca în figura 5.11, cu excepția faptului că grid-ul se extinde în afara cercului unitate.

Ecuația (5.102) arată că zerourile lui $H(z^{-1})$ sunt identice cu cele ale lui $H(z)$. Dacă $H(z)$ are un zerou complex $z = r_k \cdot e^{j\cdot\theta_k}$ atunci $H(z)$ trebuie să aibă şi o "imagine–oglindă" a acestuia, adică zeroul $z^{-1} = (1/r_k)\cdot e^{-j\cdot\theta_k}$. Pe de altă parte, dacă răspunsul la impuls este real, zerourile complexe ale lui $H(z)$ apar în perechi conjugate. Problema care apare în acest caz este menținerea proprietății de fază liniară, deoarece perechea de zerouri cuantizate $z_{3,4} = (1/r_k)\cdot e^{\pm j\cdot\theta_k}$ poate să nu fie imaginea în oglindă a perechii de zerouri cuantizate $z_{1,2} = r_k \cdot e^{\pm j\cdot\theta_k}$.

Această problemă poate fi evitată prin rearanjarea termenilor corespunzători imaginii în oglindă. Se pot scrie astfel coeficienții imaginii în oglindă, sub forma

$$\left(1 - \frac{2}{r_k}\cos\theta_k z^{-1} + \frac{1}{r_k^{\,2}}z^{-2}\right) = \frac{1}{r_k^{\,2}}\left(r_k^{\,2} - 2r_k\cos\theta_k z^{-1} + z^{-2}\right) \tag{5.108}$$

Factorul $\{1/r_k^2\}$ poate fi combinat cu câştigul total G, sau poate fi distribuit în secţiunile de filtru de ordin doi. Termenul din (5.108) conţine exact aceeaşi parametri ca şi factorul $(1 - 2r_k \cos\theta_k z^{-1} + r_k^2 z^{-2})$ şi, prin urmare, zerourile apar acum în perechi imagine-oglindă chiar dacă coeficienţii sunt cuantizaţi.

5.5. Erori cauzate de cuantizarea produselor. Caracterizarea statistică a efectelor cuantizării în realizarea în virgulă fixă a filtrelor digitale

Multiplicarea a două numere reprezentate pe b biţi fiecare, exceptând bitul de semn, are ca rezultat un număr reprezentat pe $2b$ biţi. În practică, datorită lungimii finite a registrelor cu care se lucrează, se impune exprimarea produselor prin b biţi semnificativi, astfel încât, inevitabil, cuantizarea este asociată cu formarea produsului. Indiferent de tipul de cuantizare folosit, s-a încetăţenit ca acesta să se numească *rotunjirea produsului*. Efectul acestei cuantizări asupra performanţelor filtrului depinde de modul de implementare a acestuia.

Se presupune că eroarea de rotunjire asociată formării produsului este independentă de la o iteraţie la alta, astfel încât poate fi folosit modelul cuantizării fine, sursele de zgomot fiind introduse în sistem după multiplicatoare. Astfel, multiplicatorul este modelat cu o operaţie în precizie infinită urmată de o sursă de zgomot aditiv $e[n]$, aşa încât rezultatul final să fie egal cu un nivel de cuantizare, exact cum s-a procedat la caracterizarea erorii de cuantizare la conversia A/D a unui semnal analogic.

Se începe cu caracterizarea zgomotului de rotunjire într-un filtru cauzal, cu un singur pol, care este implementat în aritmetica cu virgulă fixă şi este descris de ecuaţia neliniară cu diferenţe

$$v[n] = Q_r[av[n-1]] + x[n] \tag{5.109}$$

Efectul rotunjirii produsului $av[n\text{-}1]$ este modelat cu o secvenţă de zgomot $e[n]$ adunată la produsul necuantizat $av[n\text{-}1]$, care este

$$Q_r[av[n-1]] = av[n-1] + e[n] \tag{5.110}$$

Cu acest model pentru eroarea de cuantizare, sistemul considerat este descris de ecuaţia liniară cu diferenţe

$$v[n] = av[n-1] + x[n] + e[n] \tag{5.111}$$

Sistemul corespunzător este ilustrat în diagrama bloc din figura 5.14.

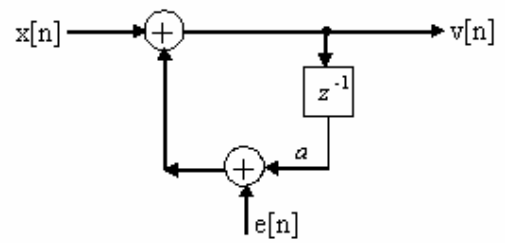

Figura 5.14. Modelul zgomotului aditiv pentru eroarea de cuantizare a produsului pentru un filtru cu un singur pol

Secvența de ieşire a filtrului $v[n]$, poate fi separată în două componente. Prima este răspunsul sistemului, $y[n]$, la secvența de intrare $x[n]$, iar a doua este răspunsul sistemului, z[n], la zgomotul aditiv de cuantizare $e[n]$. Secvența de ieşire se exprimă ca o sumă a acestor două componente, adică

$$v[n] = y[n] + z[n] \tag{5.112}$$

Înlocuind $v[n]$ din (5.112) în (5.111), se obține

$$y[n] + z[n] = ay[n-1] + az[n-1] + x[n] + e[n] \tag{5.113}$$

Pentru a simplifica analiza, se fac următoarele presupuneri în legătură cu eroarea $e[n]$:

1. Pentru orice n, secvența de eroare $\{e[n]\}$ este uniform distribuită în intervalul $\left(-\frac{1}{2}2^{-b}, \frac{1}{2}2^{-b}\right)$. Aceasta implică valoarea medie a lui $\{e[n]\}$ egală cu zero, şi dispersia
$$\sigma_e^2 = \frac{2^{-2b}}{12} \tag{5.114}$$
2. Eroarea $\{e[n]\}$ este o secvență staționară de zgomot alb şi, ca urmare, $e[n]$ şi $e[m]$ sunt necorelate pentru $n \neq m$.
3. Secvența de eroare $\{e[n]\}$ este necorelată cu semnalul $\{x[n]\}$.

Ultima presupunere permite separarea ecuației cu diferențe (5.113) în două ecuații independente:

$$y[n] = ay[n-1] + x[n] \tag{5.115}$$

$$z[n] = az[n-1] + e[n] \tag{5.116}$$

Ecuația cu diferențe (5.115) reprezintă relația de intrare-ieşire pentru sistemul dorit, iar cea din (5.116) reprezintă relația pentru eroarea de cuantizare la ieşirea sistemului.

Pentru a completa analiza se face apel la două relații importante. Prima este relația pentru valoarea medie a ieşirii $z[n]$ pentru un filtru liniar, invariant în timp, cu răspunsul la impuls $h[n]$, când este excitat de o secvență $e[n]$ cu media m_e. Rezultatul este [48]

$$m_z = m_e \sum_{n=0}^{\infty} h[n] \tag{5.117}$$

sau, echivalent,

$$m_z = m_e H(0) \tag{5.118}$$

unde $H(0)$ valoarea răspunsului în frecvență $H(\omega)$ la $\omega = 0$.

Deoarece eroarea de cuantizare datorată rotunjirii are media zero, valoarea medie a erorii la ieşire este m_z=0.

A doua relație importantă este expresia pentru secvența de autocorelație a ieşirii $z[n]$ a unui filtru cu răspunsul la impuls $h[n]$ la secvența aleatoare de intrare $e[n]$. Aceasta este [63]

$$\gamma_{zz}[n] = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h[k]h[l]\gamma_{ee}[k-l+n] \tag{5.119}$$

unde $\gamma_{ee}[n]$ este funcția de autocorelație a secvenței de intrare $e[n]$.

În cazul particular când secvența aleatoare este zgomot alb, secvența de autocorelație $\gamma_{ee}[n]$ este un impuls scalat cu dispersia σ_e^2, adică [34]

$$\gamma_{ee}[n] = \sigma_e^2 \delta[n] \tag{5.120}$$

După substituția relației (5.120) în (5.119), se obține secvența de autocorelație de la ieşirea filtrului excitat cu zgomot alb

$$\gamma_{zz}[n] = \sigma_e^2 \sum_{k=0}^{\infty} h[k]h[k+n] \tag{5.121}$$

Dispersia σ_z^2 a zgomotului de ieşire este obținută evaluând $\gamma_{zz}[n]$ la $n = 0$, adică [34]

$$\sigma_z^2 = \sigma_e^2 \sum_{k=-\infty}^{\infty} h^2[k] \tag{5.122}$$

sau, cu ajutorul teoremei lui Parseval [63], expresia alternativă

$$\sigma_z^2 = \frac{\sigma_e^2}{2\pi} \int_{-\pi}^{\pi} \left| H(\omega) \right|^2 d\omega \tag{5.123}$$

În cazul filtrului cu un singur pol, răspunsul la impuls este

$$h[n] = a^n u[n] \tag{5.124}$$

Dispersia erorii la ieşirea filtrului rezultă

$$\sigma_z^2 = \sigma_e^2 \sum_{k=0}^{\infty} a^{2k} = \frac{\sigma_e^2}{1-a^2} \tag{5.125}$$

Se observă că puterea zgomotului σ_z^2 la ieşirea filtrului este mărită față de puterea zgomotului de la intrare, σ_e^2, cu factorul $1/(1-a^2)$. Acest factor creşte odată cu apropierea polului de cercul unitate.

Fie, în continuare, un filtru recursiv de ordinul doi:

$$y[n] = -a_1 y[n-1] - a_2 y[n-2] + b_0 x[n] + b_1 x[n-1] \tag{5.126}$$

În calculul ieşirii sunt implicate patru multiplicări, dacă a_1, a_2, b_0 şi b_1 nu sunt egali cu unitatea. Zgomotul de rotunjire asociat cu fiecare multiplicare este $e_i[n]$, $i = \overline{0,3}$.

Se consideră întâi realizarea în forma directă I, ca în figura 5.15.

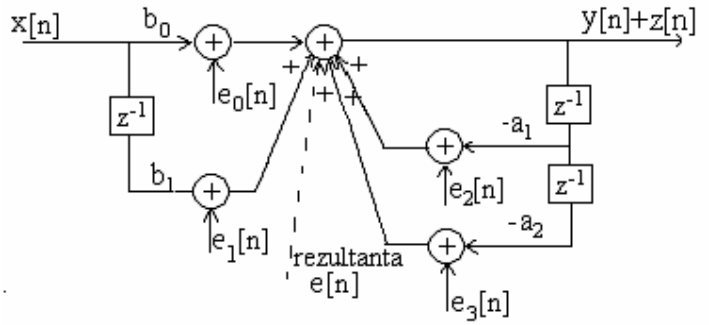

Fig. 5.15. Zgomotul de rotunjire la multiplicare pentru un filtru de ordinul doi în forma directă I

Deoarece toate sursele de zgomot se adună în acelaşi punct, acestea pot fi înlocuite cu o sursă de zgomot echivalentă

$$e[n] = \sum_{i=0}^{3} e_i[n] \tag{5.127}$$

Se observă că în implementarea în forma directă I, zgomotul trece numai prin partea de sistem ce conține numai poli, adică zerourile nu au nici un efect asupra zgomotului din ieşire.

În cazul rotunjirii, când pasul de cuantizare este constant, dispersia unei surse de zgomot este

$$\sigma_{e_i}^2 = \frac{\Delta^2}{12}, \quad i = 0, 1, 2, 3. \tag{5.128}$$

Presupunând erorile de cuantizare independente, dispersia zgomotului rezultat este suma dispersiilor componentelor

$$\sigma_{e_i}^2 = \sum_{i=0}^{3} \sigma_{e_i}^2 = \frac{\Delta^2}{3} \tag{5.129}$$

Pentru cazul general al formei directe I, când sistemul are M+1 multiplicări pentru zerouri şi N multiplicări pentru poli cu coeficienți diferiți de 0 şi 1, dispersia surselor de zgomot este

$$\sigma_e^2 = (M + N + 1)\frac{\Delta^2}{12} \tag{5.130}$$

Porțiunea din filtru prin care trece zgomotul de rotunjire este arătată în figura 5.16. Ieşirea $z[n]$ datorată zgomotului formează o parte a ieşirii cuantizate.

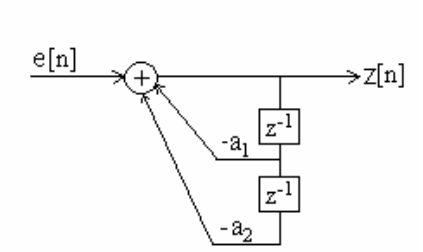

Figura 5.16. Porțiunea din filtrul recursiv afectată de zgomotul de rotunjire pentru realizarea în forma directă I.

Pentru figura 5.16 se poate scrie

$$\frac{Z(z)}{E(z)} = H'(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{5.131}$$

Evident, această funcție de transfer diferă de cea a filtrului care include şi zerouri, care este

$$H(z) = \frac{b_0 + b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{5.132}$$

Conform relației (5.74), dispersia totală de regim permanent a ieşirii datorate zgomotului de rotunjire este

$$\sigma_{0zss}^2 = \frac{\Delta^2}{3} \sum_{\substack{\text{polii din} \\ \text{cercul unitate}}} \text{reziduurile lui } H'(z) H'(z^{-1}) z^{-1} \tag{5.133}$$

cu $H'(z)$ dat de (5.131).

În cazul formei directe I dispersia totală de regim permanent a zgomotului datorat rotunjirii multiplicărilor este

$$\sigma^2_{0zss} = (M+N+1)\frac{\Delta^2}{12}\frac{1}{2\pi j}\oint H'(z)\,H'(z^{-1})z^{-1}dz = \\ = (M+N+1)\frac{\Delta^2}{12}\sum_n |h'[n]|^2 \tag{5.134}$$

unde $H'(z) = \dfrac{1}{1+\sum_{k=1}^{N} a_k z^{-k}}$ este partea care conține toți polii sistemului.

În continuare, se consideră implementarea canonică (forma directă II) a filtrului descris de (5.126), caz în care erorile de rotunjire pot fi reprezentate ca surse de zgomot poziționate ca în figura 5.17.

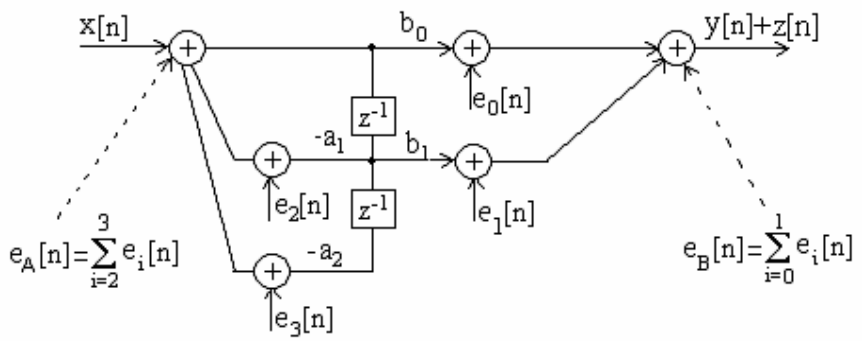

Figura 5.17. Zgomotul de rotunjire al produselor pentru un filtru recursiv implementat în forma canonică

Se observă că semnalule de eroare $e_A[n] = \sum_{i=2}^{3} e_i[n]$ cu dispersia $\dfrac{\Delta^2}{6}$, trece prin tot filtru, în timp ce $e_B[n] = \sum_{i=0}^{1} e_i[n]$ cu dispersia $\dfrac{\Delta^2}{6}$ este un zgomot adunat direct la ieşire. În acest caz dispersia de regim permanent a ieşirii datorată zgomotului de rotunjire a produselor este suma dispersiilor zgomotelor determinate de cele două semnale de eroare $e_A[n]$ şi $e_B[n]$.

$$\sigma_{0zss}^2 = \frac{\Delta^2}{6}\left[1 + \sum_{\substack{\text{polii din interiorul} \\ \text{cercului unitate}}} \text{reziduurile lui } H(z)\,H\left(z^{-1}\right)z^{-1}\right] \qquad (5.135)$$

cu $H(z)$ dat de (5.132).

Pentru cazul general al formei directe II pentru filtrul IIR, când coeficienții acestuia sunt diferiți de 0 şi 1, dispersia de regim permanent a zgomotului de ieşire este

$$\begin{aligned}\sigma_{0zss}^2 &= N\frac{\Delta^2}{12}\frac{1}{2\pi j}\oint_c H(z)\,H\left(z^{-1}\right)z^{-1}dz + (M+1)\frac{\Delta^2}{12} = \\ &= N\frac{\Delta^2}{12}\sum_n |h[n]|^2 + (M+1)\frac{\Delta^2}{12}\end{aligned} \qquad (5.136)$$

Fără a considera valori numerice pentru coeficienți, numai din compararea relațiilor (5.134) şi (5.136), nu este posibil a decide care dintre aceste forme de implementare produce un zgomot de ieşire mai mic datorat erorii de cuantizare a produselor.

Exemplul 5. 10.

Să se determine dispersia de regim permanent a zgomotului de ieşire, datorat rotunjirii aritmetice, a filtrului cu funcția de sistem

$$H(z) = \frac{b_0 + b_1 z^{-1}}{1 - 2r\cos\theta\, z^{-1} + r^2 z^{-2}}$$

implementat în

a) formă directă I

b) forma directă II

dacă r=0,9, $\theta = \pi/4$, $b_0 = 1,1$, $b_1 = 0,3$ şi pasul de cuantizare Δ.

Soluție. a) Din figura 5.15 şi 5.16 rezultă că dispersia de regim permanent a zgomotului de ieşire este

$$\sigma_{0zss}^2 = \frac{\Delta^2}{3}\sum_{\substack{\text{polii din cercul} \\ \text{unitate ai lui } H'(z)}} \text{reziduurile lui } H'(z)\,H'\left(z^{-1}\right)z^{-1} =$$

$$= \frac{\Delta^2}{3}\frac{1+r^2}{1-r^2}\frac{1}{r^4 - 2r^2\cos 2\theta + 1} = 1{,}92\,\Delta^2$$

cu $H'(z) = \dfrac{1}{1 - 2r\cos\theta\, z^{-1} + r^2 z^{-2}}$

b) Din figura (5.17) rezultă

$$\sigma^2_{0zss} = \frac{\Delta^2}{6} + \frac{\Delta^2}{6} \sum_{\text{polii lui } H(z)} \text{reziduurile lui } H(z)\,H\left(z^{-1}\right)z^{-1} =$$

$$= \frac{\Delta^2}{6}\left[1 + \frac{\left(b_0^2 + b_1^2\right)\left(1 + r^2\right) - 4b_0 b_1 r\cos\theta}{\left(r^4 - 2r^2\cos 2\theta + 1\right)\left(1 - r^2\right)}\right] = 1{,}07\,\Delta^2$$

Se observă că forma directă II (canonică) produce un zgomot de ieşire mai mic pentru valorile date ale parametrilor decât forma directa I şi că valorile b_0 şi b_1 nu afectează dispersia zgomotului de ieşire în forma directă I.

Ecuațiile (5.134) şi (5.136) arată că structurile în forma directă I şi II sunt afectate diferit de cuantizarea produselor în implementarea ecuațiilor cu diferențe corespunzătoare. În general, alte structuri echivalente, cum ar fi cele în cascadă, în paralel, lattice şi formele transpuse vor avea dispersii totale ale zgomotului la ieşire diferite de cele din structurile în formă directă. Nu se poate spune care sistem va avea dispersia de zgomot la ieşire cea mai mică, dacă nu se cunosc valorile coeficienților.

Îmbunătățirea performanței de zgomot a sistemelor numerice este posibilă folosind sumatoare şi acumulatoare pe un număr mai mare de biți. Această soluție presupune însă o complicare semnificativă a realizării "hard" a schemei.

5.6. Oscilații cu ciclu-limită în sisteme recursive

În secțiunile anterioare au fost analizate erorile care apar în operațiile aritmetice realizate de un filtru digital. Prezența unuia sau a mai multor cuantizoare în implementarea unui filtru digital, conduce la un dispozitiv neliniar a cărui caracteristică poate fi semnificativ diferită de cea a filtrului ideal. Efectele neliniare datorate aritmeticii cu precizie finită, îngreunează analiza performanțelor unui filtru digital. Pentru a efectua o analiză a efectului cuantizării, s-a adoptat o caracterizare statistică a erorilor de cuantizare, ceea ce a condus în final la un model liniar pentru filtru.

În sistemele recursive, neliniaritatea datorată efectuării operațiilor matematice în aritmetică finită poate cauza oscilații periodice la ieşire, chiar dacă secvența de intrare este zero sau o valoare constantă, nenulă. Astfel de oscilații în sistemele recursive sunt numite *cicluri limită* şi pot fi direct atribuite erorii de rotunjire sau trunchiere la multiplicare. Aceste

oscilații pot fi reduse folosind registre pe mai mulți biți. Al doilea tip de oscilații numit *oscilații de depășire* poate apărea când intrarea cuantizorului depășește domeniul dinamic. Aceste oscilații au, de obicei, amplitudine mare şi nu pot fi reduse prin creşterea numărului de biți.

5.6.1. Cicluri limită datorate rotunjirii

Fenomenul ciclurilor limită este diferit de comportamentul zgomotului cauzat de cuantizare. Efectele cuantizării se identifică cu zgomotul când nivelul semnalului este mare şi foarte variabil, făcând eroarea de cuantizare, la orice moment de timp, aproape independentă de erorile anterioare. Când nivelul semnalului este scăzut, erorile cauzate de cuantizare devin corelate. Ciclurile limită sunt periodice, dar nu neapărat sinusoidale. Ele sunt susceptibile a apărea acolo unde există reacție în filtru; filtrele IIR au întotdeauna mecanisme de reacție în interiorul lor, deci astfel de oscilații pot apărea la ieşirea lor. Spre deosebire de acestea, filtrele FIR nu conțin mecanisme de reacție şi, în consecință, ele nu vor prezenta oscilații la ieşire. Acesta este un avantaj al filtrelor FIR față de cele IIR. Tratarea generală a comportării pe cicluri limită a filtrelor digitale este dificilă, motiv pentru care se vor analiza structurile de ordinul 1 şi 2.

Pentru a ilustra caracteristica unei oscilații de ciclu limită, se consideră un sistem cu un singur pol, descris de ecuația liniară cu diferențe

$$y[n] = ay[n-1] + x[n] \tag{5.137}$$

în care polul este situat la $z=a$. Sistemul ideal este prezentat în figura 5.18a.

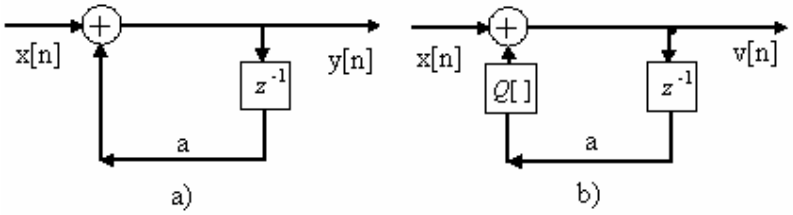

Figura 5.18. a) Sistemul recursiv ideal cu un singur pol b) Sistemul neliniar real

Sistemul real, care este descris de ecuația neliniară cu diferențe

$$v[n] = Q[av[n-1]] + x[n] \tag{5.138}$$

este realizat ca în figura 5.18b.

Se presupune că sistemul real din figura 5.18b este implementat cu o aritmetică în virgulă fixă cu patru biţi pentru amplitudine şi un bit pentru semn. Cuantizarea care se face după multiplicare este presupusă a rotunji produsul prin adaos. În Tabelul 5.4 se prezintă răspunsul sistemului real pentru patru poziţii diferite ale polului $z=a$ şi intrarea $x[n]=\beta\delta[n]$, unde β=15/16, care are reprezentarea binară 0,1111.

Tabel 5.4 Cicluri limită pentru un filtru cu un singur pol

n	a=0,1000 = 1/2	a=1,1000 = –1/2	a=0,1100 = 3/4	a=1,1000 = –3/4
0	0,1111 (15/16)	0,1111 (15/16)	0,1011 (11/16)	0,1011 (11/16)
1	0,1000 (7/16)	1,1000 (–7/16)	0,1000 (8/16)	1,1000 (–8/16)
2	0,0100 (3/16)	0,0100 (3/16)	0,0110 (6/16)	0,0110 (6/16)
3	0,0010 (1/16)	1,0010 (–1/16)	0,0101 (5/16)	1,0101 (–5/16)
4	0,0001 (1/16)	0,0001 (1/16)	0,0100 (4/16)	0,0100 (4/16)
5	0,0001 (1/16)	1,0001 (–1/16)	0,0011 (3/16)	1,0011 (–3/16)
6	0,0001 (1/16)	0,0001 (1/16)	0,0010 (2/16)	0,0010 (2/16)
7	0,0001 (1/16)	1,0001 (–1/16)	0,0010 (2/16)	1,0010 (–2/16)
8	0,0001 (1/16)	0,0001 (1/16)	0,0010 (2/16)	0,0010 (2/16)

În mod ideal, răspunsul sistemului ar trebui să scadă exponenţial spre zero ($y[n]= a^n \rightarrow 0$ când $n\rightarrow \infty$). În sistemul real, totuşi, răspunsul $v[n]$ atinge o stare stabilă periodică la ieşire, cu o perioadă ce depinde de valoarea polului. Când polul este pozitiv, oscilaţiile au loc cu perioada $N_p = 1$, astfel încât ieşirea atinge o valoare constantă de 1/16 pentru a=1/2 şi 1/8 pentru a=3/4. Acest fenomen este numit ciclu limită cu frecvenţă zero.

Pe de altă parte, când polul este negativ, secvenţa de ieşire oscilează între valori pozitive şi negative (±1/16 pentru a = –1/2 şi ±1/8 pentru a = –3/4). Prin urmare, perioada este N_p = 2. Se obţine astfel o oscilaţie de amplitudine constantă, a cărei pulsaţie este egală cu π şi a cărei amplitudine este ±1/16 sau ±1/8.

Aceste cicluri-limită apar ca rezultat al efectului de cuantizare în multiplicări. Când secvenţa de intrare $x[n]$ devine zero, ieşirea intră într-un ciclu limită după un număr de iteraţii. Ieşirea rămâne în acest ciclu limită până când este aplicat un alt semnal de intrare, suficient de puternic, pentru a scoate sistemul din ciclu. În mod similar, ciclurile limită cu intrare zero apar din condiţii iniţiale nenule. Amplitudinea ieşirii

pe perioada ciclului limită este inclusă într-un domeniu de valori care este numit "banda moartă" a filtrului. Frecvența şi amplitudinea ciclului limită depind de coeficienți, condiții inițiale, metoda de cuantizare şi lungimea cuvântului.

Este interesant de menționat faptul că atunci când răspunsul filtrului cu un pol este în ciclu limită, sistemul neliniar real lucrează ca un sistem liniar echivalent, cu un pol la z=1, atunci când polul este pozitiv (a>0), şi z = -1, când polul este negativ (a<0). Aceasta înseamnă

$$Q_r[av[n-1]] = \begin{cases} v[n-1], & a > 0 \\ -v[n-1], & a < 0 \end{cases} \tag{5.139}$$

Deoarece produsul $av[n\text{-}1]$ este rotunjit, eroarea de cuantizare este limitată de

$$\left|Q_r[a[v-1]] - av[n-1]\right| \le \frac{1}{2}2^{-b} \tag{5.140}$$

unde b este numărul de biți (exclusiv semnul) utilizat în reprezentarea polului a şi a lui $v[n]$. Prin urmare, relațiile (5.139) şi (5.140) conduc la

$$\left|v[n-1]\right| - \left|av[n-1]\right| \le \frac{1}{2}2^{-b}$$

şi, deci

$$\left|v[n-1]\right| \le \frac{\frac{1}{2}2^{-b}}{1-|a|} \tag{5.141}$$

Când coeficientul a este pozitiv, răspunsul ciclului limită se numeşte *de curent continuu* (are amplitudine şi semn constante), iar dacă a este negativ comportamentul ciclului limită are amplitudine constantă dar semn alternant.

Expresia din (5.141) defineşte *zona* sau *banda moartă* pentru un filtru cu un singur pol. De exemplu, când b = 4 şi $|a|$=1/2 banda moartă este cuprinsă în domeniul (-1/16, 1/16) pentru amplitudini, iar pentru b = 4 şi $|a|$=3/4, banda moartă creşte la (-1/8, 1/8).

Comportarea ciclului limită în cazul unui filtru cu doi poli este mult mai complexă prin faptul că poate apărea o mai mare varietate de oscilații. În acest caz sistemul ideal cu doi poli este descris de ecuația liniară cu diferențe

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + x[n] \tag{5.142}$$

în timp ce sistemul real este descris de ecuația neliniară cu diferențe

$$v[n] = Q_r[a_1 v[n-1]] + Q_r[a_2 v[n-2]] + x[n] \qquad (5.143)$$

Când coeficienții filtrului satisfac condiția $a_1^2 < -4a_2$, polii sistemului apar la $z_{1,2} = re^{\pm j\theta}$, unde $a_2 = -r^2$ şi $a_1 = 2r\cos\theta$. Ca şi în cazul filtrului cu un singur pol, când sistemul este într-un ciclu limită cu intrare zero [49],

$$Q_r[a_2 v[n-2]] = -v[n-2], \qquad (5.144)$$

adică sistemul se comportă ca un oscilator cu polii complex-conjugați situați pe cercul unitate ($a_2 = -r^2 = -1$). Rotunjirea produsului $av[n\text{-}2]$ implică

$$\left|Q_r[a_2 v[n-2]] - a_2 v[n-2]\right| \le \frac{1}{2} 2^{-b} \qquad (5.145)$$

După substituția lui (5.144) în (5.145), se obține

$$\left|v[n-2]\right| - \left|a_2 v[n-2]\right| \le \frac{1}{2} 2^{-b}$$

sau, echivalent

$$\left|v[n-2]\right| \le \frac{\frac{1}{2} 2^{-b}}{1-\left|a_2\right|} \qquad (5.146)$$

Expresia din (5.146) defineşte banda moartă a unui filtru de ordin doi cu poli complex conjugați. Se observă că limitele benzii moarte depind doar de a_2. Parametrul $a_1 = 2r\cos\theta$ determină doar frecvența oscilațiilor.

Un alt ciclu limită posibil cu intrarea zero, care este numai amintit şi care apare ca rezultat al rotunjirii multiplicărilor, corespunde unui sistem echivalent de ordinul doi cu polii la $z = \pm 1$.

Este interesant de menționat cum ciclurile limită descrise anterior au rezultat prin rotunjirea produsului dintre coeficienții filtrului şi ieşirile precedente $v[n\text{ -}1]$ şi $v[n\text{ -}2]$. În locul rotunjirii, se poate alege a trunchia produsul la b biți, caz în care se pot elimina multe din ciclurile limită, dar această soluție nu este foarte agreată, deoarece trunchierea are ca rezultat o deplasare a valorii medii a erorii, excepție făcând cazul când se foloseşte reprezentarea semn-valoare unde eroarea de trunchiere este simetrică față de zero.

În realizarea în paralel a diverselor sisteme IIR de ordin înalt cu secțiuni de ordinul doi, fiecare secțiune generează propriul ciclu limită, fără interacțiune între secțiunile de filtru de ordin doi. Prin urmare, ieşirea

este o sumă a ciclurilor limită cu intrare zero a secţiunilor individuale. În cazul realizării în cascadă pentru un sistem IIR de ordin înalt, ciclurile limită sunt mult mai greu de analizat. În particular, când prima secţiune de filtru generează un ciclu limită cu intrare zero, acesta este filtrat de secţiunile succesive. Dacă frecvenţa ciclului limită este apropiată de frecvenţa de rezonanţă a filtrului următor din succesiune, amplitudinea secvenţei va fi mărită de caracteristica de rezonanţă. În general, trebuie evitate astfel de situaţii.

5.6.2. Cicluri limită datorate depăşirii

Un tip mult mai sever de cicluri limită poate apărea datorită depăşirii aritmetice din interiorul filtrelor care folosesc aritmetica în complement faţă de unu sau în complement faţă de doi. Aceste cicluri limită sunt cunoscute sub numele de *oscilaţii de depăşire*. O depăşire la adunarea a două sau mai multe numere binare apare atunci când suma depăşeşte lungimea disponibilă a cuvântului la implementarea digitală a sistemului.

De exemplu, se consideră secţiunea de filtru de ordin doi prezentată în figura 5.19, în care adunarea se face în aritmetica complementului faţă de doi.

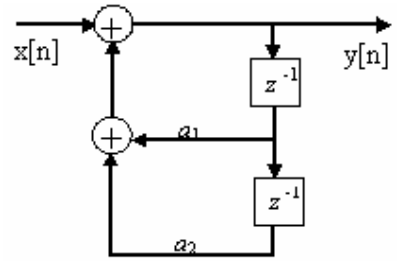

Figura 5.19. Secţiune de filtru de ordinul doi

Ieşirea din filtru se poate scrie

$$y[n] = g[a_1 y[n-1] + a_2 y[n-2] + x[n]] \qquad (5.147)$$

unde funcţia $g[.]$ reprezintă adunarea în complement faţă de doi.

Figura 5.20 prezintă caracteristica intrare-ieşire $g[v]$ a sumatorului în complement faţă de doi.

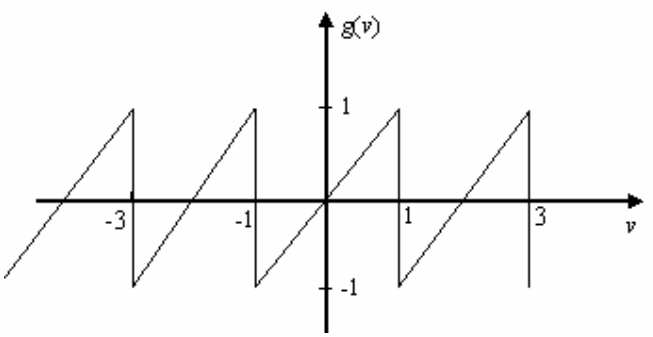

Figura 5.20. Caracteristica funcțională pentru adunarea în complement față de doi a două sau mai multe numere

Domeniul de valori al parametrilor (a_1, a_2) pentru un filtru stabil este precizat de triunghiul de stabilitate [63]. Totuşi, aceste condiții nu sunt de ajuns pentru a preveni oscilațiile datorate depăşirii din aritmetica în complement față de doi. Condiția necesară şi suficientă pentru a nu apărea cicluri limită datorate depăşirii, este [49]

$$|a_1| + |a_2| < 1 \tag{5.148}$$

care este o condiție extrem de restrictivă şi duce la o constrângere nerezonabilă asupra oricărei secțiuni de filtru de ordin doi.

Un remediu efectiv pentru rezolvarea problemei oscilațiilor provocate de depăşire este de a modifica caracteristica sumatorului, ca în figura 5.21, care operează cu saturare numerică. Atunci când este sesizată o depăşire (sau o subdepățire), ieşirea sumatorului va avea valoarea maximă de capăt de scară ±1. Distorsiunea cauzată de această neliniaritate în sumator este de obicei mică deoarece saturația apare rar. Folosirea unei astfel de neliniarități nu elimină necesitatea scalării semnalelor şi a parametrilor sistemului, aşa cum va fi descris în paragraful următor.

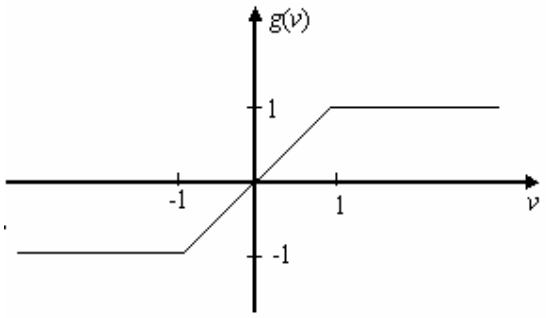

Figura 5.21. Caracteristica funcțională pentru adunare cu saturare la ±1

Ilustrarea oscilaţiilor datorate depăşirii se face pe exemplul următor. Se consideră secţiunea de filtru de ordin doi caracterizată de ecuaţia (5.143) în care adunarea se face în aritmetica complementului faţă de doi, cu lungimea cuvintelor de 4 biţi, incluzând bitul de semn, şi se foloseşte rotunjirea pentru reprezentările în complement faţă de doi. Se presupune că $a_1 = 3/4 = 0{,}110$ şi $a_2 = -3/4 = 1{,}010$ şi, de asemenea, că $x[n]$ rămâne zero pentru $n \geq 0$.

Se consideră condiţiile iniţiale $v[-1] = (3/4)_{10} = (0{,}110)_{2C}$ şi $v[-2] = (-3/4)_{10} = (1{,}010)_{2C}$. Eşantionul de la ieşire la momentul n=0 va fi

$v[0] = 0{,}110 \bullet 0{,}110 + 1{,}010 \bullet 1{,}010 = 0{,}100100 + 0{,}100100.$

Dacă se rotunjeşte fiecare produs, rezultă

$v[0] = 0{,}101 + 0{,}101 = 1{,}010 = -3/4.$

În mod similar se obţine

$v[1] = 1{,}011 + 1{,}011 = 0{,}110 = 3/4,$

adică, $v[n]$ va continua să oscileze între –3/4 şi 3/4 până ce este aplicat un semnal de intrare care să scoată sistemul din acest ciclu limită. Acesta este un exemplu de oscilaţii de depăşire. Sistemele de ordin mai mare au o comportare mai complexă.

5.7. Scalarea pentru prevenirea depăşirii

Saturaţia aritmetică descrisă în paragraful anterior elimină ciclurile limită datorate depăşirii pe de o parte, dar, pe de altă parte, duce la distorsiuni nedorite ale semnalelor, în acest caz nemaifuncţionând regula conform căreia, dacă se adună mai multe numere a căror sumă este de modul subunitar, rezultatul este corect, chiar dacă apar depăşiri în etapele intermediare de calcul. Pentru a limita aceste distorsiuni neliniare se scalează semnalul de intrare şi răspunsul la impuls între intrare şi orice nod din sistem, astfel încât să nu se depăşească gama dinamică.

Efectul depăşirii este mult mai sever pentru un filtru recursiv, decât pentru unul nerecursiv, deoarece erorile sunt filtrate din nou (datorită reacţiei) ceea ce face ca filtrul să devină inutilizabil în scurt timp. Pentru ambele tipuri de filtre, scalarea este necesară pentru reducerea amplitudinii semnalelor în anumite limite, evitându-se depăşirea în condiţii normale de lucru. Există mai multe reguli de scalare, care vor fi prezentate în cele ce urmează.

5.7.1. Norme de scalare

5.7.1.1. Scalarea după norma l_1

Se analizează toate nodurile în care ar putea apărea depăşiri şi fiecare nod din rețea este constrâns să aibă o amplitudine mai mică decât 1, pentru a evita depăşirea. Dacă $w_i[n]$ reprezintă valoarea variabilei asociată nodului *i* iar $h_i[n]$ este răspunsul la impuls de la nodul de intrare, căruia îi este asociată variabila $x[n]$, până la nodul *i,* atunci se poate scrie

$$\left|w_i[n]\right| = \left|\sum_{m=o}^{\infty} x[n-m]h_i[m]\right| . \tag{5.149}$$

Considerând că $x[n-m]$ are valoarea maximă $x_{\max}$, rezultă

$$\left|w_i[n]\right| \le x_{\max} \sum_{m=0}^{\infty} \left|h_i[m]\right| . \tag{5.150}$$

O condiție suficientă ca $\left|w_i[n]\right| < 1$ este ca

$$x_{\max} < \frac{1}{\sum_{m=0}^{\infty} \left|h_i[m]\right|} \tag{5.151}$$

pentru toate nodurile din rețea. Mărimea $l_1 = \left\|h_i\right\|_1 = \sum_{m=0}^{\infty 1} \left|h_i[m]\right|$ se numeşte *norma* l_1 a lui h_i. Dacă $x_{\max}$ nu satisface ecuația (5.151), atunci se poate multiplica $x[n]$ cu factorul de scalare $s_1 < \min_i \left\{ \frac{1}{\left\|h_i\right\|_1} \right\}$ la intrarea sistemului, astfel încât $s_1 x_{\max}$ să satisfacă (5.151) pentru toate nodurile din rețea, adică

$$s_1 x_{\max} < \frac{1}{\max_i \left[\sum_{m=0}^{\infty} \left|h_i[m]\right| \right]} \tag{5.152}$$

Scalând intrarea pe această cale se garantează că depăşirea nu apare niciodată la nici unul din nodurile de rețea. La ieşire se compensează scalarea prin înmulțirea cu $\frac{1}{s_1}$, astfel încât să nu se modifice funcția de transfer a filtrului. Relația (5.152) conduce la o

scalare foarte severă, care se mai numeşte şi *scalare de sumă*. În practică scalarea nu este făcută niciodată aşa puternic, pentru că înrăutăţeşte raportul semnal-zgomot, fapt ce va fi arătat ulterior.

5.7.1.2. Scalarea după norma l_∞

Dacă se dispune de cunoştinţe suplimentare despre intrare, se poate alege factorul de scalare, s_∞, mai mare, astfel încât să se garanteze lipsa depăşirii. Dacă intrarea este un semnal de bandă îngustă modelat cu $x[n] = x_{\max} \cos(\omega_0 n)$, variabilele de noduri vor fi [39]

$$w_i[n] = \left|H_i(\omega_0)\right| x_{\max} \cos[\omega_0 n + \angle H_i(\omega_0)] \tag{5.153}$$

Depăşirea este evitată pentru toate semnalele armonice dacă

$$\max_{i,|\omega|\le\pi} \left|H_i(\omega)\right| x_{\max} < 1 \tag{5.154}$$

Mărimea $l_\infty = \left\|H_i\right\|_\infty = \max_{|\omega|\le\pi} \left|H_i(\omega)\right|$ se numeşte norma l_∞ a lui H_i.

Dacă intrarea este scalată prin factorul de scalare $s_\infty < \min_i \left\{ \frac{1}{\left\|H_i\right\|_\infty} \right\}$ rezultă

$$s_\infty x_{\max} < \frac{1}{\max_{i,|\omega|\le\pi} \left|H_i(\omega)\right|} \tag{5.155}$$

5.7.1.3. Scalarea după norma l_2

O altă abordare posibilă este de a scala intrarea astfel încât energia fiecărei variabile de nod să fie mai mică sau egală cu energia totală a secvenţei de intrare. Se poate obţine scalarea corespunzătoare folosind inegalitatea Schwartz Buniacovski şi teorema lui Parseval [63].

$$\begin{aligned} \left|w_i[n]\right|^2 &= \left|\sum_{k=0}^{\infty} h_i[k] x[n-k]\right|^2 \le \sum_{k=0}^{\infty} \left|h_i[k]\right|^2 \sum_{k=0}^{\infty} \left|x[n-k]\right|^2 = \\ &= \sum_{k=0}^{\infty} \left|h_i[k]\right|^2 \sum_{k=0}^{\infty} \left|x[k]\right|^2 \end{aligned} \tag{5.156}$$

Pentru a asigura condiţia de nedepăşire a energiei semnalului de intrare de către varialilele de noduri, adică $\left|w_i[n]\right|^2 \le \sum_{n=0}^{\infty} \left|x[n]\right|^2$, unde

$\sum_{n=0}^{\infty}\left|x[n]\right|^2 = E_x$ este energia semnalului de intrare, se poate multiplica secvența $x[n]$ cu factorul de scalare s_2, ales astfel încât

$$s_2^2 \le \frac{1}{\max\limits_i \sum\limits_{n=0}^{\infty}\left|h_i[n]\right|^2} = \frac{1}{\max\limits_i \frac{1}{2\pi}\int\limits_{-\pi}^{\pi}\left|H_i(\omega)\right|^2 d\omega} \tag{5.157}$$

Mărimea $l_2 = \left\|h_i\right\|_2 = \left(\sum\limits_{n=0}^{\infty}\left|h_i[n]\right|^2\right)^{1/2}$ se numeşte *norma* l_2 a lui h_i.

5.7.1.4. Scalarea după norma l_p

Metodele anterioare pot fi generalizaze în sensul normei l_p.

Norma l_p a unei transformate Fourier $H(\omega)$ este definită ca [39]

$$l_p = \left\|H\right\|_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(\omega)\right|^p\right]^{\frac{1}{p}} d\omega \tag{5.158}$$

Se poate arăta că, în general, este îndeplinită inegalitatea [26]

$$\left|w_i[n]\right| \le \left\|X\right\|_p \left\|H_i\right\|_q \tag{5.159}$$

unde p şi q sunt întregi astfel încât

$$\frac{1}{p}+\frac{1}{q} = 1. \tag{5.160}$$

Pentru orice secvență $h[n]$ cu transformata Fourier $H(\omega)$ există relația [23]

$\left\|H\right\|_\infty \ge \left\|H\right\|_p$, oricare ar fi $p \in N^*$.

Ca urmare, scalarea l_∞ reduce nivelele de semnal într-o măsură mai mare decât alte scalări de tip l_p. Cele mai folosite scalări sunt l_2, l_∞, precum şi scalarea de sumă. Se poate arăta că există relația [23]

$$\sum_{n=0}^{\infty}\left|h_i[n]\right|^2 \le \max_{i,\omega}\left|H_i(\omega)\right| \le \sum_{n=0}^{\infty}\left|h_i[n]\right|, \tag{5.161}$$

adică $l_2 \le l_\infty \le l_1$.

Dintre acestea, cea mai severă este scalarea de sumă, care este şi dificil de calculat. Cel mai uşor de evaluat analitic este relația (5.157), deoarece această integrală poate fi calculată folosind teorema reziduurilor a lui Cauchy [1].

Deoarece în implementarea filtrelor recursive intervin mai multe puncte de sumare, ieşirea fiecăruia trebuie scalată pentru a evita depăşirea, deci vor fi mai multe răspunsuri la impuls $h_i[n]$ şi funcţii de sistem corespunzătoare, $H_i(z)$, care fac legătura între intrarea $x[n]$ şi semnalele intermediare $w_i[n]$.

5.7.2. Interacţiunea dintre domeniul dinamic şi zgomot

Normele de scalare l_2, l_∞, l_1 reprezintă trei moduri de a obţine coeficienţi de scalare pentru intrarea unui filtru digital. Prin scalarea intrării cu factorul $s_p, p = 1, \infty, 2$, raportul semnal / zgomot de cuantizare la ieşire scade.

În figura 22 a,b, se prezintă un sistem IIR de ordinul doi, implementat în forma directă I şi forma directă II, cu intrarea scalată. În figura 22 a, factorul de scalare s-a combinat cu coeficienţii b_k, astfel încât sursa de zgomot este aceeaşi ca în cazul fără scalare, prezentat în figura 5.15. Deoarece acest zgomot este filtrat din nou de partea de filtru care conţine polii, puterea zgomotului de ieşire este aceeaşi pentru sistemul nescalat, reprezentat în figura 5.15 şi cel scalat, reprezentat în 5.22a. Pentru sistemul din figura 22a, funcţia de sistem este $s_p H(z)$, faţă de $H(z)$ a sistemului cu intrarea nescalată şi, corespunzător, ieşirea este $y'[n] = s_p y[n]$, în loc de $y[n]$. Deoarece zgomotul este injectat după scalare, raportul dintre puterea semnalului şi cea a zgomotului în sistemul scalat este de s_p^2 ori raportul semnal/zgomot pentru sistemul nescalat din figura 5.15. Cum $s_p < 1$ atunci când este necesară scalarea, raportul semnal / zgomot la ieşirea filtrului se reduce prin scalare.

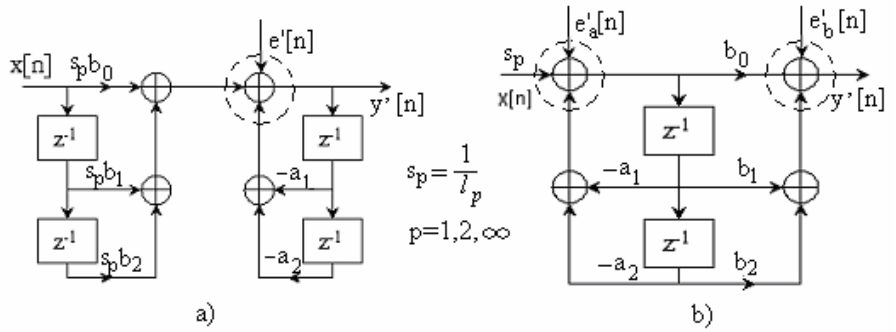

Figura 5.22. Scalarea sistemelor de ordinul doi. a) Forma directă I, b) Forma directă II

În cazul implementării în forma directă II din figura 22b factorul de scalare trebuie determinat astfel încât sa se evite depăşirea în ambele noduri încercuite. Funcția de sistem a filtrului scalat este $s_p H(z)$. Factorul de scalare s_p, $p = 1, \infty, 2$, contribuie cu o sursă suplimentară de zgomot la $e_a[n]$ a sistemului nescalat reprezentat în figura 5.17. Acest zgomot este filtrat în acelaşi mod de sistemul nescalat şi de cel scalat. Prin urmare, puterea semnalului se multiplică cu s_p^2, iar puterea zgomotului de ieşire este dată de relația (5.136), cu N înlocuit cu (N+1), astfel încât raportul semnal/zgomot se reduce şi în acest caz, dacă se efectuează scalarea pentru a evita depăşirea.

În concluzie, cu cât o regulă de scalare conduce la un factor de scalare mai scăzut, se reduce probabilitatea depăşirii, dar se reduce şi raportul semnal/zgomot de cuantizare. Acest fapt reprezintă *interacțiunea dintre domeniul dinamic şi zgomot*. Din acest motiv prezintă interes găsirea unor structuri caracterizate de zgomot de cuantizare minim în condiții de scalare precizate. Utilizarea unor structuri în formă directă de ordin mare nu conduce la rezultate satisfăcătoare din acest punct de vedere, astfel încât sunt preferate structurile în cascadă sau în paralel, realizate cu secțiuni de ordinul doi.

În continuare sunt date schemele de scalare pentru structurile în cascadă şi în paralel.

5.7.3. Scalarea în realizarea în cascadă şi în paralel

5.7.3.1. Analiza realizării în cascadă

În figura 5.23 este prezentat un sistem implementat cu K module de ordinul doi, fiecare din acestea implementat în forma canonică, conectate în cascadă.

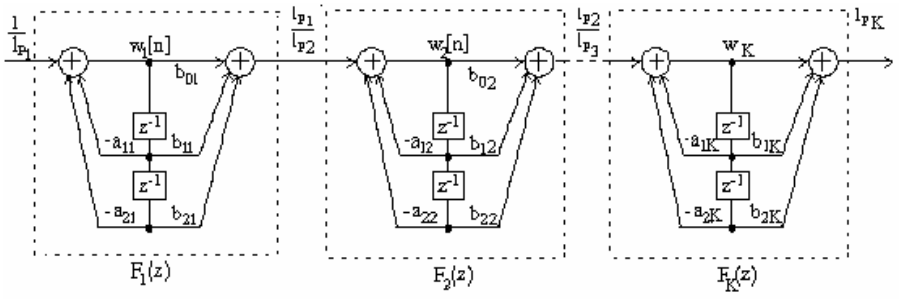

Figura 5.23. Scalarea la realizarea în cascadă a unui filtru cu K celule de ordinul doi

Se notează cu F_k, k=1,...K, funcția de sistem a unui modul de ordinul doi.

$$F_k(z) = \frac{b_{0k} + b_{1k} z^{-1} + b_{2k} z^{-2}}{1 + a_{1k} z^{-1} + a_{2k} z^{-2}} \tag{5.162}$$

$l_{pi} = \|H_i(\omega)\|_p$; i=1, 2, 3,…,K, p = 1, 2, ∞, reprezintă norma după care s-a efectuat scalarea, iar $H_i(\omega)$ - funcția de transfer de la intrare la nodul w_i.

$$H_i(z) = \frac{\prod_{k=1}^{i-1} F_k(z)}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}} \tag{5.163}$$

coeficienții $\frac{l_{pi}}{l_{p(i+1)}}$ pot fi încorporați în b_{0i}, b_{1i}, b_{2i}.

Ținând seama de cele prezentate în paragraful precedent, scalarea este propriu-zis necesară numai pentru secțiunile pentru care normele $l_{pi} = \|H_i(\omega)\|_p$ sunt supraunitare. Dacă, însă, $\|H_i(\omega)\|_p \le 1$, rezultă că nu este necesară scalare pentru celula respectivă, ceea ce ar corespunde unui factor de scalare unitar, fără efect asupra zgomotului de cuantizare. Totuşi, dacă se scalează intrarea într-o secțiune de ordinul doi cu un factor supraunitar, care va amplifica semnalul, va creşte raportul semnal/zgomot, prin utilizarea eficientă a gamei dinamice a filtrului. Astfel, scalarea poate fi privită nu numai ca un mod de a evita depăşirea, ci şi de adaptare a nivelului semnalului la gama dinamică a filtrului.

În cazul unui filtru numeric IIR de ordin mare realizat prin conectarea în cascadă a unor structuri de ordinul doi, puterea zgomotului la ieşire depinde de modul în care polii şi zerourile sunt împerecheate pentru a forma structuri de ordinul doi şi de ordinea secțiunilor în cascadă. Se poate observa că pentru *K* secțiuni de ordin doi există *K*! posibilități de a împerechea polii şi zerourile şi *K*! posibilități de a ordona secțiunile de ordinul doi rezultate. Rezultă în total $(K!)^2$ sisteme diferite. În plus, se poate alege oricare din formele directe I sau II (sau transpusele lor) pentru implementarea secțiunilor de ordinul doi. Chiar şi pentru sisteme de ordin mic problema împerecherii şi ordonării nu este simplă, deoarece necesită un volum mare de calcule.

Se defineşte *factorul* (sau *câştigul*) *de vârf* pentru celula k cu relația

$$\rho_k = \frac{\max_{\omega} |H_k(\omega)|}{\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |H_k(\omega)|^2 \, d\omega \right]^{\frac{1}{2}}} \tag{5.164}$$

În ciuda dificultății găsirii unei împerecheri şi ordonări optime, Jackson a arătat că o grupare optimă minimizează factorii de vârf şi a găsit că se pot obține rezultate bune aplicând următoarele reguli simple [23]:

1. Polul care este cel mai apropiat de cercul de rază unitate din planul Z, trebuie împerecheat cu zeroul cel mai apropiat de el;
2. Regula 1 se aplică repetat până ce toți polii şi zerourile au fost împerecheate;
3. Secțiunile de ordinul doi rezultate trebuie ordonate în funcție de apropierea polilor de cercul unitate, fie în ordinea crescătoare, fie descrescătoare a apropierii polilor de cercul unitate.

Regulile de împerechere sunt bazate pe observația că subsistemele cu câştig (factor) de vârf foarte mare sunt nedorite pentru că ele pot cauza depăşiri şi pot amplifica zgomotul de cuantizare. Împerechind un pol ce este apropiat de cercul unitate, cu un zerou adiacent se tinde să se reducă câştigul de vârf al secțiunii.

O motivație pentru regula 3 este aceea că pentru ca spectrul zgomotului de ieşire să nu aibă o alură ascuțită, cu un maxim puternic în apropierea unui pol ce este apropiat de cercul de unitate din planul *Z*, este de dorit ca aceşti poli să fie la începutul schemei în cascadă. Pe de altă parte, răspunsul în frecvență la ieşirea unui anumit nod implică produsul răspunsurilor în frecvență ale subsistemelor care preced nodul. Astfel, pentru a evita reducerea excesivă a nivelului de semnal în etajele anterioare ale cascadei ar trebui ca polii ce sunt apropiați de cercul unitate să fie plasați ultimii în cascadă. Se observă că problema ordonării secțiunilor depinde de o varietate de factori, cum ar fi dispersia totală a zgomotului de ieşire şi forma spectrului zgomotului de ieşire. Jackson a folosit norme l_p pentru a cuantifica analiza problemei împerecherii şi ordonării polilor şi zerourilor şi a elaborat o serie de reguli empirice pentru obținerea de rezultate satisfăcătoare, fără a evalua toate posibilitățile.

De multe ori, pentru obținerea unui zgomot cât mai mic, celulele se ordonează în sens crescător al factorului de vârf. În figura 5.24 este prezentată ordonarea secțiunilor de ordinul doi în cascadă în ordinea

crescătoare a selectivității, astfel încât celula cea mai selectivă să filtreze zgomotele provenite de la toate filtrele, atenuându-le.

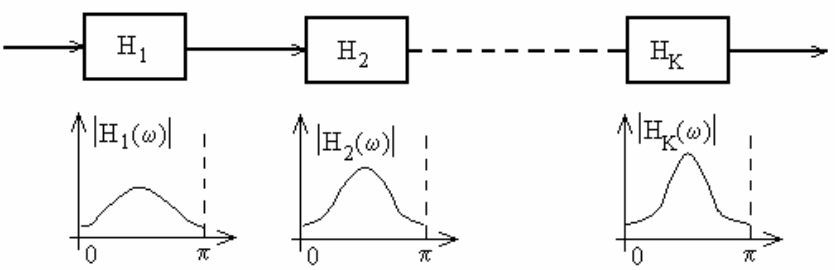

Figura 5. 24. Ordonarea secțiunilor de ordinul doi în cascadă în ordinea crescătoare a selectivității acestora

Următorul exemplu ilustrează punctul de vedere conform căruia ordonarea în cascadă a secțiunilor este importantă în controlarea zgomotului de rotunjire a produselor la ieşirea întregului sistem.

Exemplul 5.11.

Să se determine dispersia zgomotului cauzat de rotunjirea produselor, la ieşirea realizării în cascadă a filtrului cauzal, cu funcția de sistem

$$H(z) = H_1(z)H_2(z)$$

unde $$H_1(z) = \frac{1}{1-\frac{1}{2}z^{-1}}; H_2(z) = \frac{1}{1-\frac{1}{4}z^{-1}}$$

Soluție. Fie $h[n]$, $h_1[n]$, şi $h_2[n]$ răspunsurile la impuls corespunzătoare funcțiilor de transfer $H(z)$, $H_1(z)$ şi, respectiv, $H_2(z)$. Acestea sunt:

$$h_1[n] = \left(\frac{1}{2}\right)^n u[n], \quad h_2[n] = \left(\frac{1}{4}\right)^n u[n], \quad h[n] = \left[2\left(\frac{1}{2}\right)^n - \left(\frac{1}{4}\right)^n\right]u[n]$$

Cele două realizări în cascadă sunt prezentate în figura 5.25.

În prima realizare în cascadă, dispersia zgomotului la ieşire este

$$\sigma_{z1}^2 = \sigma_e^2\left[\sum_{n=0}^{\infty} h^2[n] + \sum_{n=0}^{\infty} h_2^2[n]\right]$$

În a doua realizare în cascadă, dispersia este

$$\sigma_{z2}^2 = \sigma_e^2\left[\sum_{n=0}^{\infty} h^2[n] + \sum_{n=0}^{\infty} h_1^2[n]\right]$$

$$\sum_{n=0}^{\infty} h_1^2[n] = \frac{1}{1-\frac{1}{4}} = \frac{4}{3};\ \sum_{n=0}^{\infty} h_2^2[n] = \frac{1}{1-\frac{1}{16}} = \frac{16}{15}$$

$$\sum_{n=0}^{\infty} h^2[n] = \frac{4}{1-\frac{1}{4}} - \frac{4}{1-\frac{1}{8}} + \frac{1}{1-\frac{1}{16}} = 1{,}83$$

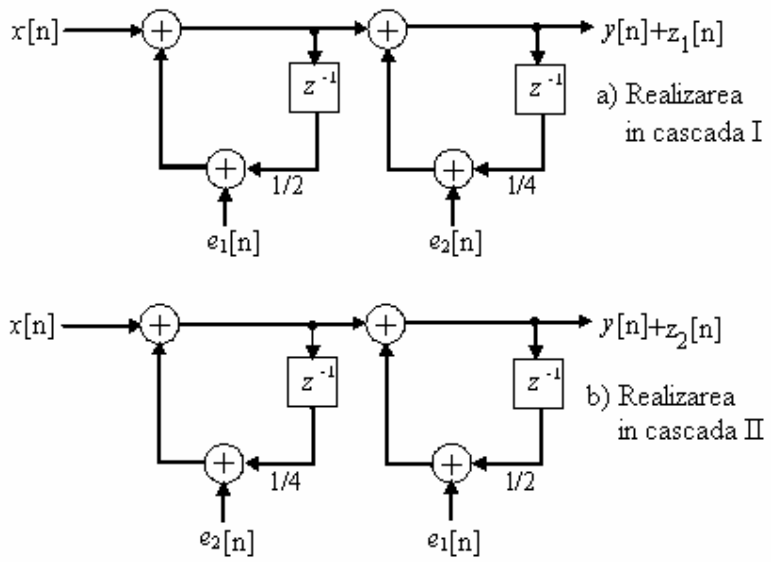

Figura 5. 25. Realizări în cascadă

În consecinţă,

$$\sigma_{z1}^2 = 2{,}90\sigma_e^2$$

$$\sigma_{z2}^2 = 3{,}16\sigma_e^2$$

iar raportul dispersiilor zgomotului de ieşire este $\frac{\sigma_{z2}^2}{\sigma_{z1}^2} = 1{,}09$.

Prin urmare, puterea zgomotului în a doua realizare în cascadă este cu 9% mai mare decât în primul caz.

5.7.3.2. Analiza realizării în paralel

În figura 5.26 este prezentat un sistem implementat cu *K* module de ordinul doi, conectate în paralel.

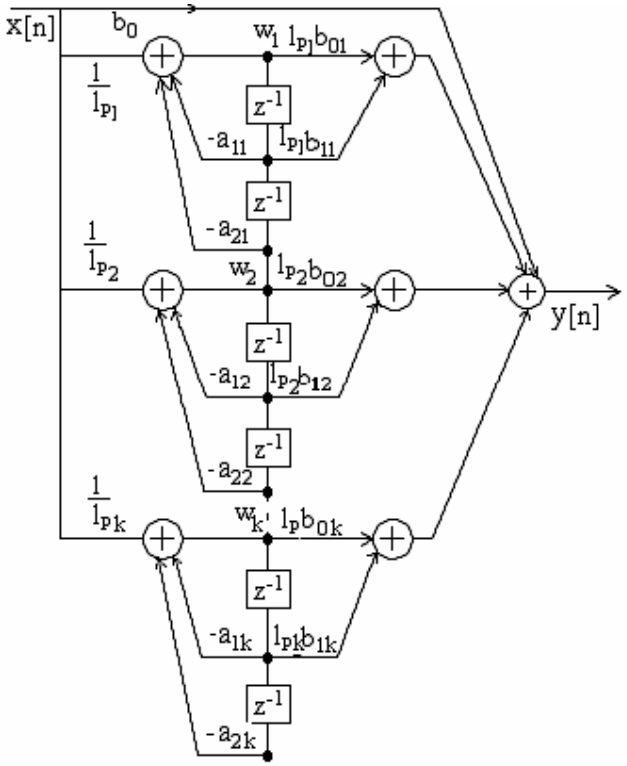

Figura 5.26. Scalarea la realizarea în paralel a unui filtru cu K celule de ordinul 2

$l_{pi} = \|H_i(\omega)\|_p$; i=1, 2, 3, …,K, p = 1, 2, ∞, reprezintă norma după care s-a efectuat scalarea.

$H_i(\omega)$ - funcția de transfer de la intrare $x[n]$ la nodul w_i. Funcția de sistem corespunzătoare este

$$H_i(z) = \frac{1}{1 + a_{1i}z^{-1} + a_{2i}z^{-2}} \text{ ; i = 1, 2, 3….} \qquad (5.165)$$

Analiza efectelor de cuantizare într-un filtru de ordin doi poate fi direct aplicată la filtrele de ordin superior bazate pe realizări în paralel. În

acest caz, fiecare secţiune de ordinul doi este independentă de celelalte secţiuni şi, deci, puterea totală a zgomotului de cuantizare la ieşire este suma puterii zgomotului de cuantizare a fiecărei secţiuni individuale. Tehnicile de împerechere enunţate anterior pot fi aplicate şi la formele în paralel unde se poate arăta [23] că puterea de zgomot la ieşire este comparabilă cu cele mai bune împerecheri şi ordonări la conectarea în cascadă. Forma în cascadă rămâne totuşi cea mai folosită pentru structurile IIR.

Deoarece structurile IIR cu formele directe I şi II includ şi sistemele FIR în forma directă ca un caz particular, rezultatele şi tehnicile de analiză considerate mai sus se aplică la sistemele FIR, dacă se elimină toate referirile la polii funcţiei de sistem şi se elimină căile de reacţie în toate grafurile de semnal.

Pentru sistemele FIR cu fază liniară, implementarea se poate face cu aproximativ jumătate din multiplicările sistemului FIR general, ceea ce determină reducerea la jumătate a dispersiei zgomotului la ieşire, dacă produsele sunt cuantizate înainte de adunare.

Rezultatele pentru realizările în cascadă de tip IIR sunt aplicabile şi pentru realizările în cascadă de tip FIR, pentru acestea urmărindu-se numai problema ordonării secţiunilor de ordinul doi.

5.7.4. Analiza erorii de cuantizare în cazul scalării intrării

Pentru a obţine o imagine mai clară a efectului erorii de cuantizare, se va considera şi efectul scalării intrării. Se reia cazul filtrului cu un singur pol din exemplul 5.7 prezentat în figura 5.12. Se presupune că secvenţa de intrare $\{x[n]\}$ este o secvenţă de zgomot alb, a cărei amplitudine a fost scalată cu norma l_1 pentru a preveni depăşirea la adunare. Atunci

$$|y[n]| \le x_{\max} \sum_{n=0}^{\infty} h[n]$$

Cum se doreşte ca $|y[n]| \le 1$, rezultă

$$x_{\max} \le \frac{1}{\sum_{n=0}^{\infty} h[n]} = 1 - |a| \qquad (5.166)$$

Dacă se presupune $x[n]$ uniform distribuit în domeniul ($-x_{\max}$, $x_{\max}$), atunci, dispersia semnalului de intrare este $\sigma_x^2 = (1- |a|)^2/3$.

Potrivit relației (5.125), puterea zgomotului la ieşirea filtrului este $\sigma_z^2 = \frac{\sigma_e^2}{1-a^2}$.

Puterea semnalului de la ieşirea filtrului este

$$\sigma_y^2 = \sigma_x^2 \sum_{k=0}^{\infty} a^{2k} = \frac{\sigma_x^2}{1-a^2} \tag{5.167}$$

Raportul dintre puterea semnalului de ieşire, σ_y^2, şi puterea erorii de cuantizare, σ_z^2, este

$$\frac{\sigma_y^2}{\sigma_z^2} = \frac{\sigma_x^2}{\sigma_e^2} = (1-|a|)^2 \cdot 2^{2(b+1)} \tag{5.168}$$

Această expresie pentru raportul semnal/zgomot de la ieşirea filtrului arată preţul plătit ca urmare a scalării intrării, mai ales când polul este apropiat de cercul unitate.

Prin comparaţie, dacă intrarea nu este scalată şi sumatorul are un număr suficient de mare de biţi pentru a evita depăşirea, amplitudinea semnalului este în intervalul (-1, 1). În acest caz, dispersia semnalului de intrare este $\sigma_x^2 = 1/3$, independentă de poziţia polului. Atunci

$$\frac{\sigma_y^{\ 2}}{\sigma_z^{\ 2}} = 2^{2(b+1)} \tag{5.169}$$

Diferenţa dintre rapoartele semnal/zgomot din (5.168) şi (5.169) demonstrează necesitatea de a utiliza mai mulţi biţi la adunare, faţă de multiplicare. Numărul biţilor adiţionali depinde de poziţia polului şi trebuie crescut odată cu mutarea polului mai aproape de cercul unitate.

În continuare, se consideră un filtru cu doi poli care, cu precizie infinită, este descris de ecuaţia liniară cu diferenţe

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + x[n] \tag{5.170}$$

unde $a_1 = 2r\cos\theta$ şi $a_2 = -r^2$.

Când cele două produse sunt rotunjite, rezultă un sistem care este descris de ecuaţia neliniară cu diferenţe

$$v[n] = Q_r[a_1 v[n-1]] + Q_r[a_2 v[n-2]] + x[n] \tag{5.171}$$

Sistemul este prezentat în schema bloc din figura 5.27.

Fiind două multiplicări, se produc două erori de cuantizare pentru fiecare ieşire.

Prin urmare, trebuie să se introducă două secvențe de zgomot e_1[n] şi e_2[n], care corespund ieşirilor cuantizoarelor

$$Q_r[a_1 v[n-1]] = a_1 v[n-1] + e_1[n]$$
$$Q_r[a_2 v[n-1]] = a_2 v[n-2] + e_2[n] \tag{5.172}$$

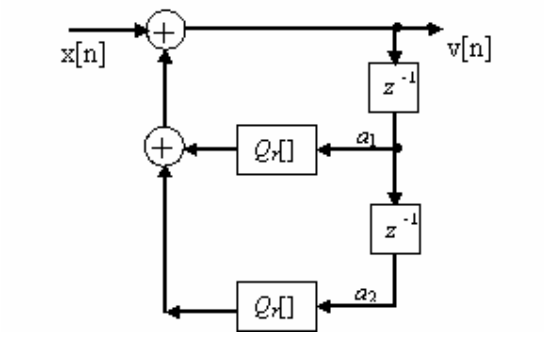

Figura 5.27 Filtru cu doi poli cu cuantizoare prin rotunjire a produselor

O diagramă bloc pentru modelul corespunzător este ilustrată în figura 5.28. Se observă că secvențele de eroare $e_1[n]$ şi $e_2[n]$ pot fi mutate direct la intrarea filtrului.

Ca şi în cazul filtrului de ordinul întâi, ieşirea filtrului de ordin doi poate fi separată în două componente, componenta semnalului dorit şi componenta erorii de cuantizare. Prima poate fi descrisă de ecuația cu diferențe

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + x[n] \tag{5.173}$$

în timp ce a doua satisface ecuația cu diferențe

$$z[n] = a_1 z[n-1] + a_2 z[n-2] + e_1[n] + e_2[n] \tag{5.174}$$

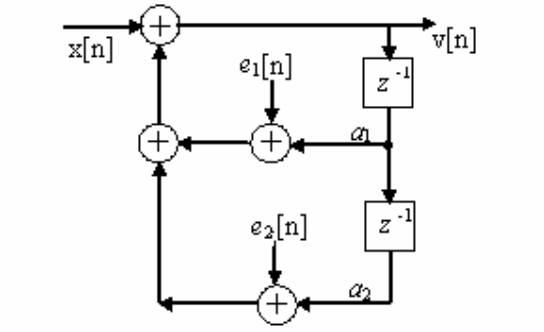

Figura 5.28 Modelul zgomotului aditiv pentru erorile de cuantizare ale unui filtru cu doi poli

Se presupune că secvențele $e_1[n]$ şi $e_2[n]$ sunt necorelate.

Răspunsul la impuls al filtrului este [63]

$$h(n) = \frac{r^n}{\sin\theta}\sin(n+1)\theta \cdot u[n] \tag{5.175}$$

Prin urmare,

$$\sum_{n=0}^{\infty} h^2[n] = \frac{1+r^2}{1-r^2}\frac{1}{r^4+1-2r^2\cos 2\theta} \tag{5.176}$$

Aplicând (5.122) se obține dispersia erorii de cuantizare la ieşirea filtrului, în forma [47]

$$\sigma_z^2 = \sigma_e^2\left(\frac{1+r^2}{1-r^2}\frac{1}{r^4+1-2r^2\cos 2\theta}\right) \tag{5.177}$$

Dacă semnalul de intrare $x[n]$ este scalat cu norma l_1 ca în (5.151) pentru a evita depăşirea, puterea semnalului de ieşire este

$$\sigma_y^2 = \sigma_x^2\sum_{n=0}^{\infty} h^2[n] \tag{5.178}$$

unde puterea semnalului de intrare $x[n]$ este dată de dispersia

$$\sigma_x^2 = \frac{1}{3\left[\sum_{n=0}^{\infty}\left|h[n]\right|\right]^2} \tag{5.179}$$

În concluzie, raportul semnal/zgomot la ieşirea filtrului cu doi poli este

$$\frac{\sigma_y^2}{\sigma_z^2} = \frac{\sigma_x^2}{\sigma_e^2} = \frac{2^{2(b+1)}}{\left[\sum_{n=0}^{\infty}\left|h[n]\right|\right]^2} \tag{5.180}$$

Cu toate că este dificilă evaluarea exactă a numitorului în (5.180), este uşor să determinăm marginile superioară şi inferioară ale acestuia. În particular, $|h[n]|$ este mărginită superior

$$\left|h[n]\right| \le \frac{1}{\sin\theta}r^n \qquad n \ge 0 \tag{5.181}$$

astfel încât

$$\sum_{n=0}^{\infty}|h[n]| \le \frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n = \frac{1}{(1-r)\sin\theta} \tag{5.182}$$

Marginea inferioară se poate obține dacă se observă că

$$|H(\omega)| = \left|\sum_{n=0}^{\infty} h[n]e^{-j\omega n}\right| \le \sum_{n=0}^{\infty}|h[n]| \tag{5.183}$$

Dar,

$$H(\omega) = \frac{1}{\left(1-re^{j\theta}e^{-j\omega}\right)\left(1-re^{-j\theta}e^{-j\omega}\right)} \tag{5.184}$$

La ω=θ, care este frecvența de rezonanță a filtrului, se obține cea mai mare valoare a lui $|H(\omega)|$, deci

$$\sum_{n=0}^{\infty}|h[n]| \ge |H(\theta)| = \frac{1}{(1-r)\sqrt{1+r^2-2r\cos 2\theta}} \tag{5.185}$$

Prin urmare, raportul semnal/zgomot este mărginit superior şi inferior conform relației

$$2^{2(b+1)}(1-r)^2\sin^2\theta \le \frac{\sigma_y^{\,2}}{\sigma_z^{\,2}} \le 2^{2(b+1)}(1-r)^2(1+r^2-2r\cos 2\theta) \tag{5.186}$$

De exemplu, când θ = π/2, expresia din (5.186) devine

$$2^{2(b+1)}(1-r)^2 \le \frac{\sigma_y^2}{\sigma_z^2} \le 2^{2(b+1)}(1-r)^2(1+r)^2 \tag{5.187}$$

Termenul dominant în aceste margini este $(1- r)^2$, care poate reduce serios raportul semnal/zgomot odată cu apropierea polilor de cercul unitate. Dacă $\delta = 1-r$ este distanța de la pol la cercul unitate, raportul semnal/zgomot din (5.187) este redus cu δ^2. Aceste rezultate servesc la întărirea aserțiuni anterioare, referitoare la necesitatea utilizării mai multor biți la adunare decât la multiplicare, ca un mecanism de evitare a erorilor rezultate din operația de scalare.