16. UTICAJ KONAČNE DUŽINE DIGITALNE REČI NA KARAKTERISTIKE SISTEMA

U dosadašnjem izlaganju prećutno je bilo pretpostavljeno da su u sistemu za obradu signala, vrednosti koeficijenata funkcije prenosa, kao i vrednosti signala, predstavljeni realnim brojevima. Ova pretpostavka važi kod sistema za obradu signala koji su diskretizovani samo po vremenu, kao što su SC filtri, CCD sistemi i SAW sistemi. Međutim, ako se realizacija sistema za obradu signala izvodi u digitalnoj formi, potrebno je izvršiti još jednu dikretizaciju signala, ovaj put po amplitudi.

Pošto se u digitalnim sistemima podaci predstavljaju samo sa konačnim brojem bita, broj amplitudskih nivoa koji se može predstaviti je ograničen. Dakle, diskretizacija po amplitudi, poznata pod nazivom kvantovanje, izaziva dodatne efekte koji u dosadašnjem proučavanju sistema za obradu signala nisu bili proučavani. Najvažniji među tim efektima su sledeći:

1. Kvantovanje koeficijenata koje ima uticaj na tačnost realizacije funkcije prenosa,
2. Kvantovanje ulaznog i izlaznog signala koje generiše šum na izlazu,
3. Kvantovanje proizvoda koje takođe generiše šum na izlazu,
4. Nelinearni efekti zbog konačne tačnosti množenja,
5. Nelinearni efekti zbog prekoračenja opsega kod sabiranja.

U daljem izlaganju prvo će ukratko biti razmotreni načini predstavljanja brojeva u digitalnim sistemima za obradu signala i analizirane greške koje se pojavljuju u raznim predstavama. Zatim će detaljnije biti razmotreni uzroci i posledice svakog od navedenih efekata u sistemima za digitalnu filtraciju signala. Na kraju će biti analiziran uticaj efekata kvantovanja na tačnost izračunavanja DFT.

16.1 PREDSTAVLJANJE BROJEVA U BINARNOM SISTEMU

U binarnom sistemu, koji se skoro isključivo koristi u savremenim računarskim sistemima, postoji nekoliko načina za predstavljanje numeričkih podataka. Pre svega, prema položaju tačke koja razdvaja celobrojni i razlomački deo broja, razlikuju se sistemi sa *fiksnom tačkom* i *pokretnom tačkom*. U okviru svakog od ovih sistema, postoji nekoliko načina predstavljanja koji se uglavnom razlikuju po načinu predstavljanja negativnih brojeva.

16.1.1 PREDSTAVLJANJE BINARNIH BROJEVA SA FIKSNOM TAČKOM

Najjednostavniji način predstavljanja pozitivnih binarnih brojeva je sistem poznat pod nazivom *prirodni binarni kod*. Pozitivni broj N se u takvom sistemu može napisati u obliku:

$$N = \sum_{i=-B}^{M} b_i 2^i = b_M b_{M-1} \ldots b_0 . b_{-1} \ldots b_{-B} \tag{16.1}$$

gde su $0 \le b_i \le 1$ cifre binarnog sistema. Krajnje levi bit binarnog broja, b_M, se naziva *bit najveće težine* (engl. most significant bit - MSB), dok se krajnje desni bit, b_{-B}, naziva *bit najmanje težine* (engl. least significant bit - LSB). Težinski faktor uz LSB, 2^{-B}, predstavlja razliku dva susedna broja, odnosno *rezoluciju* binarne predstave. U praksi se najčešće, zbog određenih pogodnosti, sistemom sa fiksnom tačkom predstavljaju brojevi koji predstavljaju prave razlomke, tj. čiji je celobrojni deo jednak nuli.

Da bi se realizovalo predstavljanje negativnih brojeva, potrebno je uvesti još jedan bit koji će pokazivati znak broja. Binarni kodovi za predstavljanje brojeva sa znakom (označenih brojeva) nazivaju se *bipolarni kodovi*. Danas se koriste četiri bipolarna koda: *znak plus amplituda*, *pomereni binarni kod*, *komplement jedinice* i *komplement dvojke*.

Najprostiji bipolarni binarni kod je *znak plus amplituda* (engl. sign-magnitude). Krajnje levi bit predstavlja bit znaka i ima vrednost 0 za pozitivne brojeve i vrednost 1 za negativne brojeve. Preostalih B bitova predstavljaju apsolutnu vrednost (amplitudu) broja. Nepogodnosti ovog sistema su dvostruka reprezentacija nule (00...0 ili 10...0) i komplikovano izvođenje sabiranja, pa se zbog toga relativno retko koristi u praksi.

Pomereni binarni kod se izvodi iz prirodnog binarnog koda. Naime, najnegativnijem broju se dodeljuje vrednost 00...0 a najpozitivnijem vrednost 11...1. Dakle, pomereni binarni kod je ekvivalentan prirodnom binarnom kodu brojeva dobijenih sabiranjem sa konstantom, koja je jednaka apsolutnoj vrednosti najnegativnijeg broja. Nedostatak ovog načina predstavljanja je komplikovano izvođenje računskih operacija. Još jedan nedostatak ovog koda je što je vrednost MSB, 0 za negativne, a 1 za pozitivne brojeve, što je u suprotnosti sa uobičajenom konvencijom. Prednost ovog načina predstavljanja je u jednostavnijoj realizaciji A/D i D/A konvertora.

Treći bipolarni kod predstavlja *komplement jedinice* ili *prvi komplement*. Pozitivni brojevi se predstavljaju prirodnim binarnim kodom kod kojih je MSB jednak 0. Negativni brojevi dobijaju se komplementiranjem cifara predstave apsolutne vrednosti broja (nule se zamene jedinicama i obratno). Nedostaci ovog metoda su dvostruka reprezentacija nule (00...0 i 11...1) i komplikovanije sabiranje, jer se eventualni prenos na MSB poziciji mora sabrati sa LSB cifrom u rezultatu.

Četvrti način predstavljanja negativnih brojeva u binarnom obliku je *komplement dvojke* ili *drugi komplement*. Na ovaj način, predstava negativnog broja se dobija oduzimanjem pozitivnog broja od 2. Jednostavniji način je da se prvo komplementiranjem cifara pozitivnog broja formira komplement jedinice, a da se posle toga dobijenom broju doda 1 LSB. Prednost komplementa dvojke nad ostalim sistemima je što se najjednostavnije izvodi operacija sabiranja, koja je najčešća u računarskim sistemima, jer se sabiranje izvodi na uobičajeni način, a eventualni prenos na poziciji MSB se ignoriše. Druga prednost je jedinstvena predstava nule, što olakšava testiranje rezultata. Treća prednost, koju ima samo ovaj sistem predstave, je što je *rezultat sabiranja više brojeva korektan ako leži u dozvoljenom opsegu* $-1 \le N \le 1 - 2^{-B}$, *čak i ako neki međurezultati leže izvan dozvoljenog opsega*. S obzirom na navedene prednosti, *komplement dvojke predstavlja najčešće korišćeni način predstavljanja označenih brojeva u savremenim računarskim sistemima*, pa i u većini sistema za digitalnu obradu signala. Vrednost binarnog broja u kodu komplementa dvojke ($b_0 . b_{-1} b_{-2} \ldots b_{-B}$) data je izrazom sličnim sa (16.1), gde bit b_0 predstavlja znak broja.

$$N = -b_0 + \sum_{i=1}^{B} b_{-i} 2^{-i} \tag{16.2}$$

U Tabeli 16.1 su prikazani gore opisani bipolarni kodovi sa četiri bita, od kojih jedan predstavlja znak. Iz Tabele 16.1 se može uočiti jedna interesantna osobina: pomereni binarni kod i komplement dvojke imaju komplementarne bitove znaka, dok su ostali bitovi identični. Ova osobina olakšava sintezu A/D i D/A konvertora koji koriste komplement dvojke.

Tabela 16.1 Četvorobitni bipolarni kodovi.

	Znak+amplituda	Pomereni binarni kod	Komplement 1	Komplement 2
0.875	0111	1111	0111	0111
0.750	0110	1110	0110	0110
0.625	0101	1101	0101	0101
0.500	0100	1100	0100	0100
0.375	0011	1011	0011	0011
0.250	0010	1010	0010	0010
0.125	0001	1001	0001	0001
0.000	0000/1000	1000	0000/1111	0000
-0.125	1001	0111	1110	1111
-0.250	1010	0110	1101	1110
-0.375	1011	0101	1100	1101
-0.500	1100	0100	1011	1100
-0.625	1101	0011	1010	1011
-0.750	1110	0010	1001	1010
-0.875	1111	0001	1000	1001
-1.000	-	0000	-	1000

U digitalnoj obradi signala je uobičajeno da se svi signali i svi koeficijenti predstavljaju brojevima koji su *pravi razlomci* sa $B+1$ bita, gde je $B+1$ najčešće 16, 24 ili 32. Pri množenju dva takva broja dobija se rezultat koji takođe predstavlja pravi razlomak i ima $2B+1$ važećih bita tako da se mora skratiti na polaznu dužinu. U slučaju sabiranja dva prava razlomka od $B+1$ bita može doći do *prekoračenja opsega* (engl. overflow), odnosno rezultat može biti veći od najpozitivnijeg dozvoljenog broja ili manji od najnegativnijeg dozvoljenog broja. U tom slučaju moraju se predvideti dodatne mere radi korektne interpretacije rezultata.

16.1.2 PREDSTAVLJANJE BINARNIH BROJEVA SA POKRETNOM TAČKOM

Osnovni nedostatak predstave brojeva sa fiksnom tačkom je što je rezolucija konstantna i određena brojem upotrebljenih bitova za predstavljanje broja. Ako je potrebno povećati opseg brojeva izvan $-1 \le N < 1$, bez povećanja broja bita, to se može izvesti samo pomeranjem položaja tačke udesno, što izaziva smanjenje rezolucije.

Predstavljanje brojeva sa pokretnom tačkom uvedeno je radi povećanja dinamičkog opsega. Broj N se u sistemu sa pokretnom tačkom predstavlja u obliku:

$$N = M \cdot 2^p \tag{16.3}$$

gde je M *mantisa*, označeni broj sa fiksnom tačkom iz opsega $0.5 \le M < 1$, dok je p označeni celi broj koji se naziva *eksponent* ili *karakteristika*.

Raspoloživi broj bita za predstavljanje broja se može na razne načine raspodeliti na mantisu i karakteristiku. Na primer, ako je na raspolaganju 32 bita, uobičajeno se za mantisu koristi $23+1=24$ bita, dok se za karakteristiku koristi $7+1=8$ bita. Interesantno je uporediti dinamički opseg i rezoluciju koji se može ostvariti ovakvom raspodelom bitova na mantisu i karakteristiku sa dinamičkim opsegom i rezolucijom koji se ostvaruje sa istim brojem bita u reprezentaciji sa fiksnom tačkom. Ako se u reprezentaciji sa fiksnom tačkom koristi 32 bita, od kojih je jedan za znak, onda je rezolucija 2^{-31}, dok je najveći broj koji se može predstaviti $1-2^{-31}$. Dinamički opseg je približno 2^{31}. U opisanoj reprezentaciji sa pokretnom tačkom, najbolja rezolucija je $0.5\cdot 2^{-128}\approx 1.5\cdot 10^{-39}$, dok je maksimalni broj koji se može predstaviti $(1-2^{-23})\cdot 2^{127}\approx 1.7\cdot 10^{38}$. Dinamički opseg je približno $1.16\cdot 10^{77}$, ali sa promenljivom rezolucijom, koja je finija za male brojeve i grublja za veće brojeve.

Kod sabiranja brojeva sa pokretnom tačkom potrebno je prvo pomeriti cifre mantise manjeg broja udesno, dok se karakteristike oba broja ne izjednače, zatim se izvodi sabiranje mantisa i na kraju svođenje mantise u dozvoljeni opseg uz korekciju karakteristike (*normalizacija mantise*). Množenje dva broja u predstavi sa pokretnom tačkom vrši se tako što se pomnože mantise, saberu karakteristike i na kraju izvrši normalizacija mantise.

Predstavljanje brojeva sa pokretnom tačkom poboljšava dinamički opseg ne narušavajući rezoluciju i u tom pogledu znatno nadmašuje sisteme predstavljanja sa fiksnom tačkom. Osnovni nedostaci sistema sa pokretnom tačkom leže u povećanoj složenosti hardvera za izvođenje aritmetičkih operacija kao i u manjoj brzini izračunavanja. Razlog za oba nedostatka je jasan; prilikom svake aritmetičke operacije moraju se vršiti operacije i sa mantisom i sa karakteristikom što produžava vreme izvršenja operacije ili složenost hardvera za 30-50%. Zbog svojih prednosti, predstavljanje brojeva sa pokretnom tačkom se koristi u svim primenama gde brzina obrade nije važna, kao što je to slučaj u obradi signala van realnog vremena na računarima opšte namene.

16.1.3 PREDSTAVLJANJE BINARNIH BROJEVA U BLOKOVIMA SA POKRETNOM TAČKOM

U praksi se ponekad se koristi još jedan sistem za predstavljanje binarnih brojeva koji predstavlja hibrid između sistema sa fiksnom i pokretnom tačkom. U ovakvom načinu predstavljanja, grupe (blokovi) brojeva su predstavljeni u obliku (16.3) ali imaju fiksni eksponent p. Zajednički eksponent za grupu brojeva se određuje tako što se ispitaju svi brojevi u grupi, a zatim se za eksponent p izabere eksponent koji odgovara predstavi najvećeg broja iz grupe u sistemu sa pokretnom tačkom. To znači da mantise ostalih brojeva mogu biti i manje od 0.5. Prednost ovakvog sistema je u uštedi memorije kod hardverskih realizacija, kao i u jednostavnijem izvođenju aritmetičkih operacija. Ovaj način predstavljanja se najviše koristi u hardverskoj realizaciji DFT algoritama, ali se može koristiti i u drugim primenama gde se radi sa podacima koji su složeni u vektore ili matrice.

16.1.4 GREŠKE ZBOG ODSECANJA I ZAOKRUŽIVANJA BINARNIH BROJEVA

Prilikom izvođenja računskih operacija u sistemima za digitalnu obradu signala često se javlja potreba da se binarni broj skrati, tj. predstavi manjim brojem bita. Skraćivanje (kvantovanje) binarnog broja vrši se korišćenjem dva postupka: *odsecanja* i *zaokruživanja*. Pošto se prilikom

skraćivanja reprezentacije broja neminovno čini neka greška, interesantno je ispitati osobine grešaka koje se tom prilikom pojavljuju.

Posmatrajmo broj sa fiksnom tačkom x koji je pre skraćivanja bio predstavljen sa $L+1$ bitom i koji treba predstaviti sa $B+1$ bitom, gde je $B<L$. Ako $Q(x)$ predstavlja operaciju kvantovanja, greška kvantovanja se može definisati izrazom:

$$\varepsilon = Q(x) - x \tag{16.4}$$

Osobine greške ε zavise od sistema predstavljanja binarnog broja i metoda kvantovanja što će biti predmet analize u ovom odeljku. Izvedeni rezultati važiće i za kvantovanje analognog signala prilikom A/D konverzije ako se stavi $L \to \infty$.

Posmatrajmo prvo slučaj odsecanja, kod koga se jednostavno svi bitovi na pozicijama $B+1,\ldots,L$ odbacuju. Razmotrićemo slučajeve predstavljanja pomoću znaka i amplitude i komplementa dvojke koji su najčešći u praksi. Ako je broj koji se skraćuje pozitivan, u oba načina predstavljanja dobija se identična reprezentacija. Pošto se odsecanjem dobija broj koji je sigurno manji od broja koji je skraćen (osim u slučaju kada su svi odbačeni bitovi jednaki 0), za grešku odsecanja važi nejednakost:

$$-(2^{-B} - 2^{-L}) \le \varepsilon_T \le 0, \quad x \ge 0 \tag{16.5}$$

gde se maksimalna greška dobija kada su svi odbačeni bitovi jednaki 1.

Kvantovanjem negativnog broja u sistemu znak plus amplituda dobija se broj manje apsolutne vrednosti, tako da je greška odsecanja pozitivna i ograničena nejednakostima:

$$0 \le \varepsilon_T \le 2^{-B} - 2^{-L}, \quad x < 0 \tag{16.6}$$

Kvantovanjem negativnog broja predstavljenog komplementom dvojke dobija se:

$$\varepsilon = \left(-b_0 + \sum_{i=1}^{B} b_{-i} 2^{-i} \right) - \left(-b_0 + \sum_{i=1}^{L} b_{-i} 2^{-i} \right) = -\sum_{i=B+1}^{L} b_{-i} 2^{-i} \tag{16.7}$$

odnosno,

$$-(2^{-B} - 2^{-L}) \le \varepsilon_T \le 0, \quad x < 0 \tag{16.8}$$

Dakle, *u sistemu znak plus amplituda greška odsecanja leži u opsegu koji je simetričan oko nule*:

$$-(2^{-B} - 2^{-L}) \le \varepsilon_T \le 2^{-B} - 2^{-L} \tag{16.9}$$

dok je *u sistemu komplementa dvojke greška odsecanja uvek negativna* i leži u opsegu:

$$-(2^{-B} - 2^{-L}) \le \varepsilon_T \le 0 \tag{16.10}$$

Skraćivanje reprezentacije broja zaokružavanjem vrši se tako da se suvišni bitovi odbacuju ako je bit na poziciji $B+1$ jednak 0, dok se u slučaju kada je bit na poziciji $B+1$ jednak 1 izvrši sabiranje rezultata odsecanja sa jedinicom na poziciji B. U praksi se zaokruživanje jednostavnije izvodi dodavanjem jedinice na poziciji $B+1$ i prostim odsecanjem. Greška zaokruživanja može biti i pozitivna i negativna, a maksimalna vrednost apsolutne greške iznosi $0.5(2^{-B} - 2^{-L})$. Dakle, *greška zaokruživanja je simetrična oko nule i leži u opsegu*:

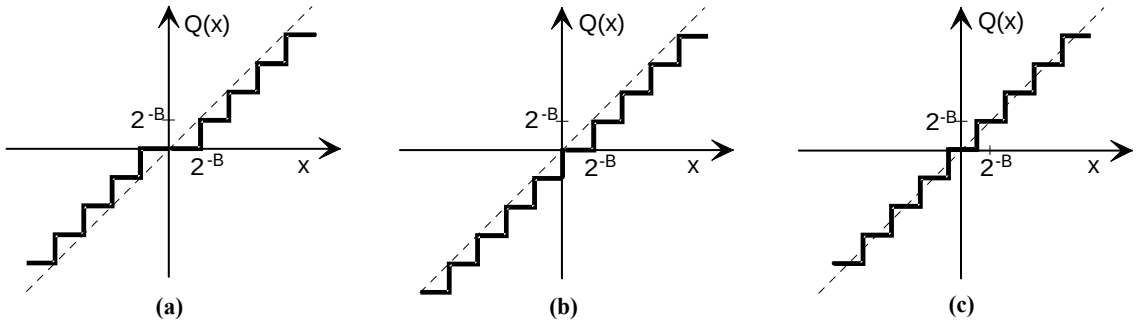
$$-0.5(2^{-B} - 2^{-L}) \le \varepsilon_R \le 0.5(2^{-B} - 2^{-L}) \tag{16.11}$$

Na slici 16.1 su prikazane greške kvantovanja za tri analizirana slučaja i $L \to \infty$.

Ako je broj x predstavljen sa pokretnom tačkom, onda se proces kvantovanja primenjuje na mantisu. Situacija je tada komplikovanija jer je greška proporcionalna vrednosti broja zbog toga što je rezolucija promenljiva. Za razliku od (16.4), grešku kvantovanja broja predstavljenog sa pokretnom tačkom pogodno je predstaviti u obliku:

$$\varepsilon = \frac{Q(x) - x}{x} \tag{16.12}$$

tj., ε *predstavlja relativnu grešku*.



Slika 16.1 Greške kvantovanja: (a) odsecanje, znak plus amplituda, (b) odsecanje, komplement dvojke, (c) zaokruživanje.

U slučaju kvantovanja mantise pozitivnog broja $2^{p-1} \le x < 2^p$ odsecanjem, na apsolutnu grešku εx se može primeniti prethodno izvedeni zaključak (16.5), čime se dobija:

$$-2^p 2^{-B} < \varepsilon_T x < 0, \quad x \ge 0 \tag{16.13}$$

odakle se posle deljenja sa 2^{p-1} dobija:

$$-2^{-B+1} < \varepsilon_T \le 0, \quad x \ge 0 \tag{16.14}$$

Ako je negativni broj predstavljen komplementom dvojke, onda iz (16.10) i (16.12) sledi:

$$-2^p 2^{-B} \le \varepsilon_T x \le 0, \quad x < 0 \tag{16.15}$$

odnosno,

$$0 \le \varepsilon_T < 2^{-B+1}, \quad x < 0 \tag{16.16}$$

Ako se vrši zaokruživanje mantise, onda je apsolutna greška simetrična oko nule i maksimalno iznosi $\pm 2^{-B-1}$. Dakle,

$$-2^p 2^{-B-1} < \varepsilon_R x \le 2^p 2^{-B-1} \tag{16.17}$$

odnosno, posle deljenja sa 2^{p-1}:

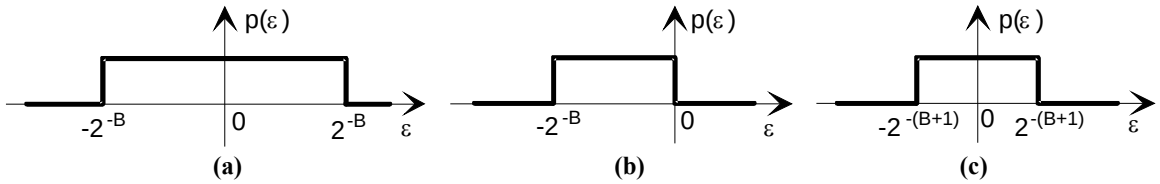
$$-2^{-B} < \varepsilon_R \le 2^{-B} \tag{16.18}$$

U analizi uticaja grešaka kvantovanja na rad sistema za digitalnu obradu signala uobičajeno se koristi statistički pristup. U tom slučaju se proces kvantovanja nekog broja predstavlja kao dodavanje *aditivnog šuma* nekvantovanoj vrednosti, odnosno iz (16.4) sledi:

$$Q(x) = x + \varepsilon \tag{16.19}$$

Greška kvantovanja ε se obično modeluje kao slučajna promenljiva, koja leži u opsegu koji zavisi od načina predstavljanja binarnog broja i načina kvantovanja. Takođe se obično

podrazumeva da slučajna promenljiva ε ima uniformnu raspodelu. Gustine verovatnoća grešaka kvantovanja u tri analizirana slučaja kvantovanja brojeva sa fiksnom tačkom su prikazane na slici 16.2.



Slika 16.2 Gustina verovatnoće greške kvantovanja: (a) odsecanje, znak plus amplituda, (b) odsecanje, komplement dvojke, (c) zaokruživanje.

16.2 KVANTOVANJE ULAZNOG I IZLAZNOG SIGNALA

U sistemima za digitalnu obradu signala, na ulazu sistema se nalazi A/D konvertor za konverziju analognog u digitalni signal, dok se na izlazu nalazi D/A konvertor za konverziju digitalnog u analogni signal. Uticaj oba konvertora na modifikaciju frekvencijskih karakteristika sistema je analiziran u trećem poglavlju. Međutim, pored uticaja na frekvencijsku karakteristiku, pošto oba konvertora za predstavljanje digitalnih informacija koriste konačni broj bita, pojavljuju se greške u reprezentaciji signala koje se nazivaju *greške kvantovanja* i koje će biti predmet proučavanja u ovom odeljku.

16.2.1 KVANTOVANJE ULAZNOG SIGNALA

Realni A/D konvertor u sistemima za digitalnu obradu signala ima dvostruku ulogu: da obezbedi odabiranje kontinualnog signala konstantnom i dovoljno velikom učestanošću odabiranja i da izvrši kvantovanje ulaznog signala na potreban broj nivoa koji se predstavljaju u binarnoj formi. Dakle, na izlazu A/D konvertora se dobija diskretizovana i kvantovana predstava analognog signala:

$$x_q[n] = Q\big(x[n]\big) = Q\big[x_a(nT)\big] \tag{16.20}$$

Proces kvantovanja se najčešće vrši postupkom zaokruživanja, tj. vrednosti analognog signala se pridružuje najbliži kvantizacioni nivo. Kriva kvantovanja ulaznog signala je prikazana na sl. 16.3 i veoma je slična sa krivom kvantovanja sa zaokruživanjem sa slike 16.1c. Jedina razlika je u prirodi promenljive x. Na slici 16.1, x je digitalna veličina, a na slici 16.3, x je kontinualna veličina. Vidi se da je proces kvantovanja nelinearan i neinvertibilan, jer se više vrednosti ulaznog signala preslikavaju u istu vrednost izlaznog signala. Greška kvantovanja može se predstaviti sekvencom:

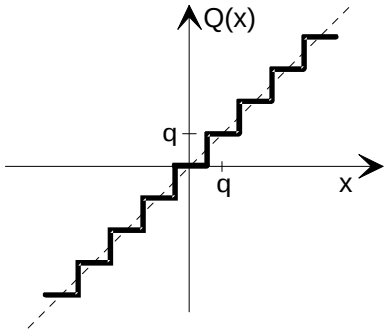
$$\varepsilon[n] = x_q[n] - x[n] \tag{16.21}$$

Neka je signal $x[n]$ normalizovan tako da leži u opsegu $-1 \le x[n] < 1$. Ako se u postupku kvantovanja koristi $B+1$ bit onda je *rezolucija* ili *korak kvantovanja*:

$$q = \frac{1}{2^B} = 2^{-B} \tag{16.22}$$

a greška kvantovanja leži u opsegu:

$$-\frac{q}{2} < \varepsilon[n] < \frac{q}{2} \tag{16.23}$$



Slika 16.3 Kriva kvantovanja ulaznog signala sa zaokružavanjem na najbližu vrednost.

Ispitivanje uticaja greške kvantovanja se najjednostavnije može izvršiti ako se pretpostavi da je $\varepsilon[n]$ sekvenca slučajnih brojeva, koja zadovoljava sledeće uslove:

1. Greška kvantovanja $\varepsilon[n]$ ima uniformnu gustinu raspodele u opsegu datom sa (16.23),
2. Greška kvantovanja $\varepsilon[n]$ predstavlja stacionarni beli šum,
3. Greška kvantovanja $\varepsilon[n]$ nije korelisana sa signalom $x[n]$.

U opštem slučaju, ova tri uslova nisu zadovoljena. Na primer, osobina 2 zahteva da greška $\varepsilon[n]$ i greška $\varepsilon[m]$ za $m \neq n$ nisu korelisane, što nije zadovoljeno ako je nivo signala mali. Međutim, u velikom broju praktičnih situacija, kada je korak kvantovanja mali, tako da se dva sukcesivna odbirka signala $x[n]$ razlikuju za nekoliko nivoa, sva tri navedena uslova su zadovoljena pa se *greška kvantovanja može smatrati za aditivni beli šum* koji se dodaje signalu. Uticaj aditivnog šuma se obično izražava preko *odnosa signal/šum* (*SNR*) koji se definiše kao:

$$SNR\,(\text{dB}) = 10\log\frac{P_x}{P_n} \tag{16.24}$$

gde je P_x snaga ulaznog signala a P_n snaga šuma kvantovanja.

Ako je gustina raspodele greške kvantovanja uniformna u intervalu (16.23), onda je srednja vrednost greške jednaka nuli, a snaga šuma kvantovanja (varijansa) je data izrazom:

$$P_n = \sigma_\varepsilon^2 = \int_{-q/2}^{q/2} \varepsilon^2 p(\varepsilon) d \qquad \varepsilon = \frac{1}{q}\int_{-q/2}^{q/2} \varepsilon^2 d \qquad \varepsilon = \frac{q^2}{12} = \frac{2^{-2B}}{12} \tag{16.25}$$

tako da je odnos signal/šum:

$$SNR\,(\text{dB}) = 10\log\frac{P_x}{P_n} = 10\log P_x + 10\log(12\cdot 2^{2B}) = 10\log P_x + 10.8 + 6.02B \tag{16.26}$$

Dakle, *svaki dodatni bit u A/D konverziji ulaznog signala povećava odnos signal/šum kvantovanja za 6 dB*, odnosno, smanjuje snagu šuma kvantovanja za 6 dB. Takođe se vidi da povećanje snage ulaznog signala za dva puta povećava odnos signal/šum za 6 dB. Iz toga proizilazi zaključak da amplituda ulaznog signala treba da bude maksimalno velika, onoliko koliko dozvoljavaju ulazne karakteristike A/D konvertora. Ako je napon pune skale A/D konvertora $X_{\max} \neq 1$, onda se izraz (16.26) usložnjava i postaje:

$$SNR\,(\text{dB}) = 10\log P_x - 10\log X_{\max}^2 + 10.8 + 6.02B = 20\log\frac{\sigma_x}{X_{\max}} + 10.8 + 6.02B \qquad (16.27)$$

U veoma čestom slučaju, kada je analogni ulazni signal govor ili muzika, raspodela amplituda je vrlo slična Gausovoj raspodeli. Srednja vrednost signala je nula, a verovatnoća da amplituda signala bude tri do četiri puta veća od srednje kvadratne (efektivne) vrednosti signala σ_x je vrlo mala. Ako se pretpostavi da je raspodela amplituda Gausova, onda je verovatnoća da amplituda odbirka bude veća od $4\sigma_x$ samo 0.00064. Da bi se izbeglo prekoračenje opsega A/D konvertora, obično se usvaja $X_{\max} = 4\sigma_x$, tako da se iz (16.27) dobija:

$$SNR\,(\text{dB}) = 6.02B - 1.25 \qquad (16.28)$$

Šum koji se postupkom kvantovanja unese u ulazni signal se ne može eliminisati i pojavljuje se na izlazu. Zbog uticaja frekvencijske karakteristike sistema za obradu signala, snaga šuma na izlazu koji potiče od šuma kvantovanja se može izračunati prema izrazu (12.55), dok se gustina spektra snage izlaznog šuma može izračunati prema (12.57).

16.2.2 KVANTOVANJE IZLAZNOG SIGNALA

Ovaj izvor šuma se vrlo često zanemaruje, mada može biti značajan. Naime, da bi se smanjio uticaj kvantovanja koeficijenata, kao i šuma koji se generiše prilikom množenja, a koji će biti detaljnije analizirani u narednim odeljcima, koeficijenti množača i signali se predstavljaju sa većom tačnošću od tačnosti ulaznog ili izlaznog signala, a aritmetičke operacije se izvode sa još većom tačnošću. A/D i D/A konvertori koji se koriste u digitalnoj obradi signala najčešće imaju rezoluciju 8-16 bita, dok se interne aritmetičke operacije u savremenim digitalnim procesorima signala izvode sa tačnošću od 16-64 bita. Dakle, pre uvođenja signala u D/A konvertor potrebno je izvršiti predstavljanje izlaznog signala manjim brojem bita. Time se uvodi dodatni izvor šuma koji povećava nivo izlaznog šuma za oko 3 dB.

16.3 KVANTOVANJE KOEFICIJENATA

U postupku sinteze funkcija prenosa IIR ili FIR tipa najčešće se koriste računarski programi napisani u nekom višem programskom jeziku. U takvim programima promenljive su obično predstavljene u binarnom sistemu sa pokretnom tačkom, pri čemu se najčešće koristi dužina reči od 32 bita (23+1 bit za mantisu i 8 bita za eksponent). Kako se u postupku realizacije sistema za digitalnu obradu signala, na osnovu izabrane realizacione strukture, najčešće koristi specijalizovani hardver, koji koristi znatno manji broj bita, pojavljuje se problem uticaja predstavljanja koeficijenata na osobine sistema. Naravno, ostvarene vrednosti koeficijenata će se razlikovati od vrednosti koeficijenata određenih postupkom sinteze. Zbog toga će se pojaviti pomeraj polova i nula funkcije prenosa u odnosu na idealni položaj, a takođe će doći do deformacija amplitudske i fazne karakteristike. Za projektanta sistema je od velikog interesa da proceni kolike su te deformacije, kako bi odabrao optimalan broj bita za reprezentaciju signala i koeficijenata. Ova procena ne mora da bude suviše tačna, jer je kod savremenih digitalnih sistema broj bita obično multipl broja 4, ili još češće broja 8. Tačno određivanje broja bita potrebno je samo u retkim slučajevima kada se projektuje kompletan hardver za određenu namenu u VLSI tehnologiji.

16.3.1 KVANTOVANJE KOEFICIJENATA KOD IIR SISTEMA

Kod IIR sistema, kvantovanje koeficijenata utiče na promenu položaja i nula i polova funkcije prenosa. Neka je funkcija prenosa $H(z)$, dobijena nekim od postupaka sinteze iz poglavlja 8, prikazana na uobičajeni način kao količnik dva polinoma:

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}} \tag{16.29}$$

gde su koeficijenti a_k i b_k prikazani sa dovoljno velikim (idealno beskonačnim) brojem bita. Koeficijenti a_k i b_k ujedno predstavljaju i koeficijente množača u direktnim realizacijama funkcije prenosa koje su prikazane na slikama 7.11, 7.12, 7.13 i 7.14. Ako se izvrši kvantovanje koeficijenata, dobija se nova funkcija prenosa:

$$\hat{H}(z) = \frac{\sum_{k=0}^{M} \hat{b}_k z^{-k}}{1 + \sum_{k=1}^{N} \hat{a}_k z^{-k}} \tag{16.30}$$

gde su $\hat{a}_k = a_k + \Delta a_k$ i $\hat{b}_k = b_k + \Delta b_k$ nove vrednosti koeficijenata dobijene kvantovanjem.

16.3.1.1 Uticaj kvantovanja koeficijenata na položaj polova funkcije prenosa

Neka su polovi funkcije prenosa sa tačnim vrednostima koeficijenata u tačkama $z = p_i$, $i = 1, \ldots, N$. Radi jednostavnosti izvođenja, pretpostavimo da su polovi prosti. Tada se imenilac funkcije prenosa $H(z)$ može predstaviti u obliku:

$$D(z) = 1 + \sum_{k=1}^{N} a_k z^{-k} = \prod_{i=1}^{N} (1 - p_i z^{-1}) \tag{16.31}$$

Polovi funkcije prenosa sa kvantovanim koeficijentima leže u tačkama $z = p_i + \Delta p_i$, $i = 1, \ldots, N$. Pomeraj i-tog pola može se izraziti u funkciji grešaka koeficijenata totalnim diferencijalom:

$$\Delta p_i = \sum_{k=1}^{N} \frac{\partial p_i}{\partial a_k} \Delta a_k, \quad i = 1, \ldots, N \tag{16.32}$$

Kako je:

$$\left(\frac{\partial D(z)}{\partial a_k} \right)_{z=p_i} = \left(\frac{\partial D(z)}{\partial p_i} \right)_{z=p_i} \frac{\partial p_i}{\partial a_k}, \quad i = 1, \ldots, N, \; k = 1, \ldots, N \tag{16.33}$$

iz (16.31) se dobija:

$$\frac{\partial p_i}{\partial a_k} = \frac{-p_i^{N-k}}{\prod_{\substack{j=1 \\ j \neq i}}^{N} (p_i - p_j)}, \quad i = 1, \ldots, N, \; k = 1, \ldots, N \tag{16.34}$$

Zamenom izračunatih parcijalnih izvoda (16.34) u (16.32), može se odrediti pomeraj i-tog pola Δp_i zbog kvantovanja koeficijenata polinoma u imeniocu funkcije prenosa. Potpuno analogni

rezultat može se izvesti za pomeraje nula funkcije prenosa, koji zavise isključivo od kvantovanja koeficijenata polinoma u brojiocu funkcije prenosa.

Posmatranjem izraza (16.34) može se uočiti jedna interesantna činjenica. Ako su polovi funkcije prenosa bliski (što je karakteristično za vrlo selektivne funkcije prenosa), onda *male greške koeficijenata polinoma u imeniocu mogu izazvati velike pomeraje polova u direktnim realizacionim strukturama*. Sa porastom broja polova u grupi, osetljivost polova na promene koeficijenata raste. Iz tog razloga se, ako je broj polova veći od 3, umesto direktnih struktura koriste kaskadne ili paralelne realizacione strukture, opisane u odeljcima 7.3.3 i 7.3.4. Kod takvih struktura se svaki par konjugovano kompleksnih polova realizuje nezavisno od ostalih polova. Zbog toga je pomeraj pola zbog greške kvantovanja koeficijenata nezavisan od položaja ostalih polova funkcije prenosa. Kod kaskadne realizacije ista osobina važi i za nule funkcije prenosa.

Kod paralelne realizacije, pomeraj nula zavisi od kvantovanja svih koeficijenata polinoma u brojiocu i imeniocu funkcije prenosa. I pored toga, pokazuje se da je paralelna realizacija ipak bolja od direktnih realizacija u pogledu osetljivosti na kvantovanje koeficijenata zbog male osetljivosti pojedinačnih sekcija drugog reda.

Dakle, *kaskadna struktura je manje osetljiva na kvantovanje koeficijenata od paralelne realizacije, a obe su znatno bolje od direktnih realizacija*.

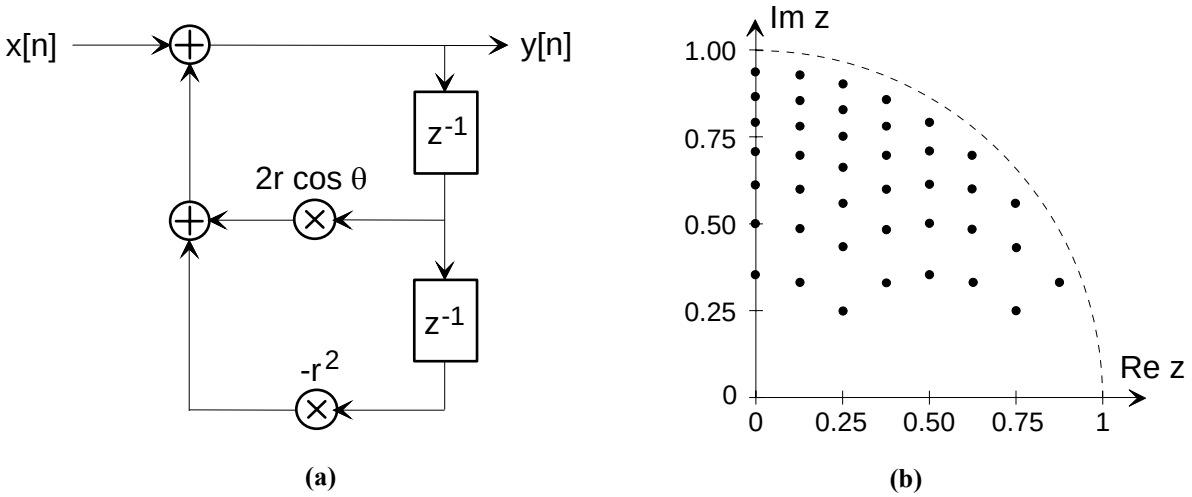
Osim kaskadne i paralelne realizacije, malu osetljivost na kvantovanje koeficijenata imaju i rešetkaste realizacije, koje su takođe opisane u sedmom poglavlju. U literaturi su opisane i neke druge realizacione strukture sa manjom osetljivošću na kvantovanje koeficijenata, kao što su na primer, talasni digitalni filtri (engl. wave digital filters), lestvičaste realizacije, realizacije zasnovane na digitalnoj verziji generalisanog konvertora impedanse, itd. Međutim, ove strukture su manje aktuelne s obzirom da su znatno složenije, a za dužine reči iznad 16 bita nemaju značajnijih prednosti nad kaskadnom i paralelnom realizacijom.

I pored male osetljivosti sekcija drugog reda koje se koriste u kaskadnoj ili paralelnoj realizaciji, postoji mogućnost da se njihova osetljivost još više smanji. Neka je, na primer, konjugovano kompleksni par polova, $z = re^{\pm j\theta}$, realizovan direktnom strukturom sa slike 16.4a kojoj odgovara funkcija prenosa:

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} = \frac{1}{1 - (2r\cos\theta) z^{-1} + r^2 z^{-2}} \tag{16.35}$$

Zbog kvantovanja koeficijenata $2r\cos\theta$ i $-r^2$ polovi mogu da zauzmu samo konačan broj položaja u z ravni koji zavisi od broja bita kojima se predstavljaju koeficijenti. Mogući položaji kompleksnih polova nalaze se na preseku koncentričnih krugova (zbog kvantovanja $-r^2$) i vertikalnih linija (zbog kvantovanja $2r\cos\theta$). Na slici 16.4b predstavljene su moguće lokacije polova u prvom kvadrantu ako su koeficijenti predstavljeni sa 4 bita, od čega je jedan iskorišćen za znak.

U slučaju predstavljanja sa 4 bita, r^2 može imati 7 različitih pozitivnih vrednosti, dok $2r\cos\theta$ može imati 7 pozitivnih vrednosti, 8 negativnih vrednosti i nulu. Na osnovu položaja polova u prvom kvadrantu, može se lako odrediti i položaj polova u ostala tri kvadranta, simetričnim preslikavanjem u odnosu na realnu i imaginarnu osu. Sa slike 16.4b se može uočiti da su mogući položaji polova neravnomerno raspoređeni u z ravni i da ih ima ukupno 146. U okolini realne ose postoji vrlo mali broj mogućih položaja, tako da polovi kod kojih je $\theta \approx 0$, ili $\theta \approx \pi$, mogu imati veliki pomeraj zbog kvantovanja koeficijenata.

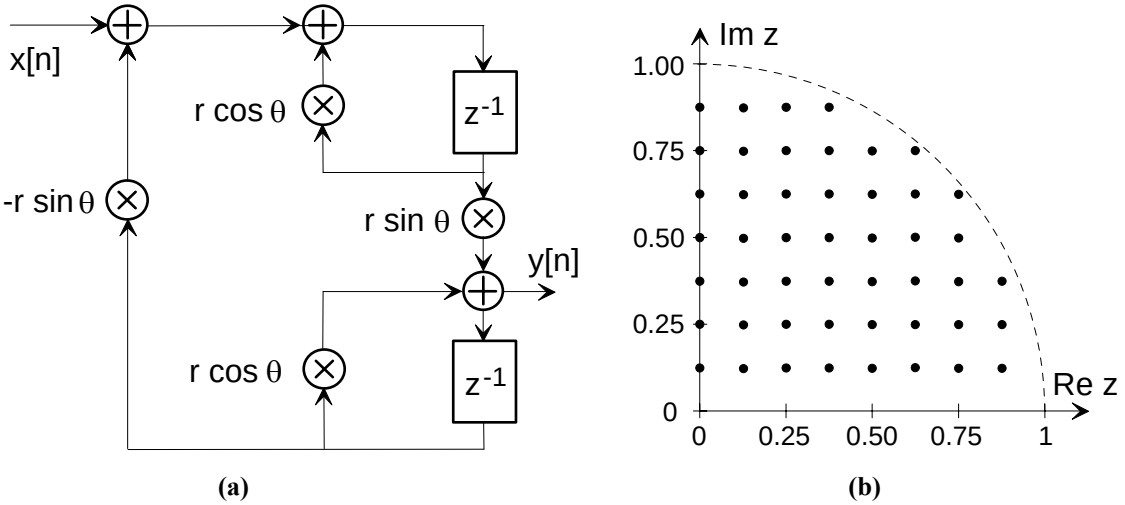


(a) **(b)**

Slika 16.4 (a) Direktna realizacija konjugovano kompleksnog para polova, (b) Moguće lokacije polova u prvom kvadrantu za predstavljanje koeficijenata sa 3+1 bita.

Zbog činjenice da i sekcije drugog reda u nekim slučajevima mogu imati veliku osetljivost na kvantovanje koeficijenata, razvijen je veći broj alternativnih realizacija ćelija drugog reda. Međutim, veću primenu našla je samo realizacija poznata pod nazivom *spregnuta forma*, koja je prikazana na slici 16.5a, a čija je funkcija prenosa data izrazom:

$$H(z) = \frac{r \sin\theta z^{-1}}{1-(2r\cos\theta)z^{-1}+r^2 z^{-2}} \tag{16.36}$$



(a) **(b)**

Slika 16.5 (a) Spregnuta realizacija konjugovano kompleksnog para polova, (b) Moguće lokacije polova u prvom kvadrantu za predstavljanje koeficijenata sa 3+1 bita.

Spregnuta struktura sa slike 16.5a ima iste kompleksne polove, $z = re^{\pm j\theta}$, kao i direktna struktura sa slike 16.4a, ako se koristi beskonačno veliki broj bita za predstavljanje binarnih brojeva. Međutim, ako se koristi konačan broj bita, u spregnutoj strukturi se kvantuju koeficijenti množača, $r\cos\theta$ i $\pm r\sin\theta$, koji ujedno predstavljaju realni i imaginarni deo polova. Zbog toga mogući položaji polova leže na presecima ekvidistantnih horizontalnih i vertikalnih linija u z ravni. Na slici 16.5b prikazani su mogući položaji položaji polova u prvom kvadrantu za predstavljanje koeficijenata množača sa 3+1 bita. Očigledno je da ovakav raspored mogućih položaja ima znatne prednosti nad rasporedom sa slike 16.4b, jer su mogući položaji ravnomernije raspoređeni u kompleksnoj ravni, a ukupan broj mogućih položaja je povećan na 178. Međutim, spregnuta

struktura ima 4 množača, tj. znatno je komplikovanija i skuplja za realizaciju. Ipak, u nekim slučajevima, kada sa kratkom dužinom reči treba ostvariti malu grešku u položaju polova, ovakva struktura ima korisnu primenu.

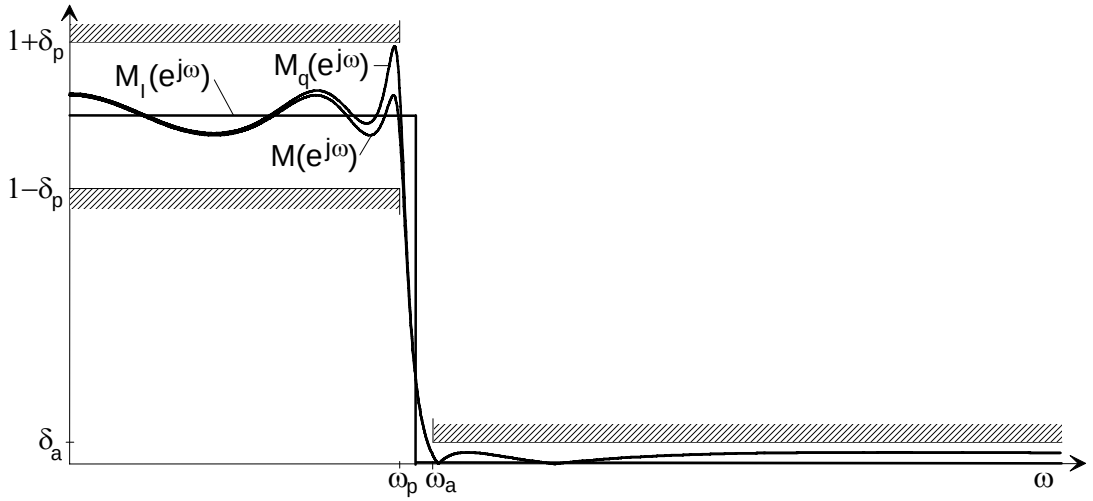
16.3.1.2 Određivanje potrebnog broja bita za realizaciju funkcije prenosa

U prethodnom odeljku analiziran je uticaj kvantovanja koeficijenata množača na pomeraj polova i nula funkcije prenosa. Međutim, za praksu je mnogo značajnija deformacija amplitudske karakteristike koja nastaje zbog kvantovanja koeficijenata, zbog toga što se gabariti najčešće zadaju za amplitudsku karakteristiku. Neka je $M(e^{j\Omega})$ amplitudska karakteristika funkcije prenosa pre kvantovanja koeficijenata, a $M_q(e^{j\Omega})$ amplitudska karakteristika posle kvantovanja koeficijenata. Neka je, takođe, $M_I(e^{j\Omega})$ amplitudska karakteristika idealnog filtra, koja se u postupku sinteze aproksimira sa dozvoljenom greškom δ_p u propusnom opsegu i δ_a u nepropusnom opsegu. Na slici 16.6 su prikazani gabariti koji se zadaju pri projektovanju NF funkcije prenosa kao i grafici funkcija $M(e^{j\Omega})$, $M_q(e^{j\Omega})$ i $M_I(e^{j\Omega})$. Zbog kvantovanja koeficijenata pojavljuje se izobličenje amplitudske karakteristike:

$$\Delta M(\Omega) = M(e^{j\Omega}) - M_q(e^{j\Omega}) \tag{16.37}$$

tako da amplitudska karakteristika može izaći izvan postavljenih gabarita. Ako se maksimalna dozvoljena vrednost za izobličenje amplitudske karakteristike usled kvantovanja označi sa $\Delta M_{\max}(\Omega)$, onda se prema slici 16.6 može pisati:

$$\Delta M_{\max}(\Omega) = \begin{cases} \delta_p - \left| M(e^{j\Omega}) - M_I(e^{j\Omega}) \right| & \Omega \le \Omega_p \\ \delta_a - \left| M(e^{j\Omega}) - M_I(e^{j\Omega}) \right| & \Omega \ge \Omega_a \end{cases} \tag{16.38}$$



Slika 16.6 Uticaj kvantovanja koeficijenata na funkciju prenosa.

Tražene specifikacije biće zadovoljene posle kvantovanja koeficijenata ako je uslov:

$$\Delta M(\Omega) \le \Delta M_{\max}(\Omega) \tag{16.39}$$

zadovoljen na svim učestanostima. Dakle, može se izvršiti izračunavanje $\Delta M(\Omega)$ za neku dužinu reči i ispitivati da li je zadovoljena nejednakost (16.39). *Optimalna dužina reči je minimalna dužina reči za koju važi uslov* (16.39). Ovim postupkom se potrebna dužina reči tačno određuje, ali je sprovođenje postupka zametno i zahteva programsku simulaciju kvantovanja koeficijenata.

Kako u praksi obično nije potrebno poznavati dužinu reči sa tačnošću od 1 bit, naročito s obzirom na činjenicu da se najčešće koriste dužine reči od 16, 24 ili 32 bita, razvijeni su alternativni postupci za određivanje optimalne dužine reči, koji su zasnovani na statističkim principima.

Postupak određivanja statistički optimalne dužine reči biće objašnjen na primeru sistema koji koristi aritmetiku sa fiksnom tačkom i kod koga se kvantovanje koeficijenata vrši zaokruživanjem. Odstupanja koeficijenata Δc_i u takvom sistemu predstavljaju slučajne brojeve, koji se nalaze u opsegu $-q/2 \le \Delta c_i \le q/2$ i imaju uniformnu raspodelu verovatnoće sa nultom srednjom vrednošću. Varijansa odstupanja koeficijenata je:

$$\sigma^2_{\Delta c_i} = \int_{-q/2}^{q/2} (\Delta c_i)^2 \, p(\Delta c_i) d(\Delta c_i) = \frac{q^2}{12} = \frac{2^{-2B}}{12} \tag{16.40}$$

Odstupanje amplitudske karakteristike, $\Delta M(\Omega)$, takođe je slučajna veličina, jer zavisi od odstupanja koeficijenata Δc_i. Prema razvoju u Tejlorov red imamo:

$$\Delta M(\Omega) = \sum_{i=1}^{N_c} \frac{\partial M(\Omega)}{\partial c_i} \Delta c_i = \sum_{i=1}^{N_c} S_{c_i} \Delta c_i \tag{16.41}$$

gde je S_{c_i} osetljivost amplitudske karakteristike $M(\omega)$ na promenu koeficijenta c_i. Za slučajnu veličinu $\Delta M(\Omega)$, pod pretpostavkom da su odstupanja koeficijenata statistički nezavisna, važi:

$$E\{\Delta M\} = \sum_{i=1}^{N_c} S_{c_i} E\{\Delta c_i\} = 0 \tag{16.42}$$

$$\sigma^2_{\Delta M} = \sum_{i=1}^{N_c} S^2_{\Delta c_i} \sigma^2_{\Delta c_i} = \frac{q^2}{12} \sum_{i=1}^{N_c} S^2_{\Delta c_i} = \frac{q^2}{12} S_T^2 \tag{16.43}$$

Pošto je broj koeficijenata, N_c, najčešće veliki, na osnovu centralne granične teoreme sledi da $\Delta M(\omega)$ ima Gausovu raspodelu, odnosno:

$$p(\Delta M) = \frac{1}{\sigma_{\Delta M}\sqrt{2\pi}} e^{-\frac{(\Delta M)^2}{2\sigma^2_{\Delta M}}} \tag{16.44}$$

Ako se sa y označi verovatnoća da se odstupanje amplitude $\Delta M(\Omega)$ nalazi u simetričnom opsegu $-\Delta M_1 \le \Delta M(\Omega) \le \Delta M_1$, onda je:

$$y = \Pr(|\Delta M| \le \Delta M_1) = \frac{2}{\sigma_{\Delta M}\sqrt{2\pi}} \int_0^{\Delta M_1} e^{-\frac{(\Delta M)^2}{2\sigma^2_{\Delta M}}} d(\Delta M) \tag{16.45}$$

Ako se uvedu smene:

$$\Delta M = x\sigma_{\Delta M} \text{ i } \Delta M_1 = x_1\sigma_{\Delta M} \tag{16.46}$$

jednačina (16.45) se može dovesti na standardni oblik:

$$y = \frac{2}{\sqrt{2\pi}} \int_0^{x_1} e^{-\frac{x^2}{2}} dx \tag{16.47}$$

Nažalost, integral u (16.47) se ne može rešiti u zatvorenom obliku. Međutim, zbog toga što se jednačina (16.47) često pojavljuje u rešavanju problema u fizici i tehnici, na raspolaganju su opsežne tabele i numerički metodi, kojima se za izabranu vrednost *y* može odrediti granica integrala x_1, odnosno ΔM_1. Dakle, dobijena *granica* ΔM_1 *predstavlja statističku granicu za* ΔM *sa faktorom pouzdanosti y*. Tada je i uslov

$$\Delta M_1 \le \Delta M_{\max}(\Omega) \tag{16.48}$$

zadovoljen sa faktorom pouzdanosti *y*, a potrebna dužina reči se naziva *statistička dužina reči*. Iz jednačina (16.43), (16.46) i (16.48) se dobija optimalni korak kvantovanja:

$$q \le \frac{\sqrt{12}\Delta M_{\max}(\Omega)}{x_1 S_T} \tag{16.49}$$

odakle sledi da je potrebni broj bita desno od tačke:

$$B = \log_2 \frac{1}{q} = \log_2 \frac{x_1 S_T}{\sqrt{12}\Delta M_{\max}(\Omega)} \tag{16.50}$$

Ako koeficijenti nisu normalizovani na opseg $-1 \le c_i < 1$, onda je za predstavljanje koeficijenta, levo od tačke, pored jednog bita za znak potrebno još dodatnih *M* bita:

$$M = \log_2 \left[\max_{1 \le i \le N_c} |c_i| \right] \tag{16.51}$$

pa je ukupan potrebni broj bita (statistička dužina reči):

$$WL = 1 + B + M \tag{16.52}$$

Prema literaturi [C-41] i [C-42], odlično slaganje između tačne i statističke dužine reči se dobija ako se izabere $x_1 = 2$, čemu odgovara faktor pouzdanosti 0.95.

Statistička procena dužine reči predstavlja pogodnu meru za procenu osetljivosti raznih realizacionih struktura na kvantovanje koeficijenata. Ponekad se statistička dužina reči koristi i kao kriterijumska funkcija u optimizacionim algoritmima za minimizaciju dužine reči.

16.3.1.3 Optimizacija kvantovanih koeficijenata

U prethodnom odeljku izvršena je statistička analiza uticaja kvantovanja koeficijenata, kojom se može prilično tačno odrediti potreban broj bita za reprezentaciju koeficijenata za zadatu dozvoljenu grešku amplitudske karakteristike. Međutim, u nekim praktično značajnim slučajevima, potrebno je sa zadatom dužinom reči ostvariti što tačniju amplitudsku karakteristiku. U takvim slučajevima se ne vrši prosto skraćivanje izračunatih tačnih vrednosti koeficijenata, već se pribegava postupku optimizacije koeficijenata.

U literaturi [A-13] razvijen je optimizacioni postupak, kojim se podešavaju vrednosti koeficijenata minimizacijom kriterijumske funkcije:

$$\max_{\omega} E(e^{j\Omega}) = \max_{\omega} \frac{H_I(e^{j\Omega}) - \hat{H}(e^{j\Omega})}{\delta(\Omega)} \tag{16.53}$$

gde je $H_I(e^{j\Omega})$ funkcija prenosa idealnog filtra, koja, na primer, u slučaju propusnika niskih učestanosti ima oblik:

$$H_I(e^{j\Omega}) = \begin{cases} 1 & 0 \le \Omega \le \Omega_p \\ 0 & \Omega_a \le \Omega \le \pi \end{cases} \tag{16.54}$$

dok je $\delta(\Omega)$ težinska funkcija oblika:

$$\delta(\Omega) = \begin{cases} \delta_p & 0 \le \Omega \le \Omega_p \\ \delta_a & \Omega_a \le \Omega \le \pi \end{cases} \tag{16.55}$$

Postupak optimizacije se izvodi u diskretnom parametarskom prostoru, odnosno, koeficijenti mogu uzimati samo određene diskretne vrednosti koje se mogu realizovati sa zadatim brojem bita. Optimizacija se izvodi tako što se, za izabrani skup koeficijenata ispituje $\max E(e^{j\Omega})$. Ako je $\max E(e^{j\Omega}) > 1$ rešenje nije prihvatljivo, pa se koriguju vrednosti koeficijenata. Optimizacioni postupak se prekida kada se dobije prihvatljivo rešenje, tj. kada je $\max E(e^{j\Omega}) \le 1$. Kao početno rešenje za optimizacioni postupak koriste se koeficijenti dobijeni prostim kvantovanjem.

U literaturi [A-13] naveden je primer kvantovanja koeficijenata eliptičkog filtra propusnika opsega osmog reda. Kada su koeficijenti skraćivani bez optimizacije, dobijena je minimalna dužina reči od 11 bita. Međutim, kada je, posle skraćivanja koeficijenata, sproveden optimizacioni postupak, isti filtar je bilo moguće realizovati sa dužinom reči od samo osam bita. Drugi eksperimenti su pokazali da se znatno smanjenje potrebne dužine reči može dobiti povećanjem reda funkcije prenosa iznad minimalnog u postupku aproksimacije. Osim toga, na izbor dužine reči sa kojom će se realizovati dati sistem utiču, osim tačnosti karakteristika, i drugi faktori kao što su na primer, složenost realizacije, brzina rada, generisani šum, ekonomičnost realizacije, itd.

16.3.2 KVANTOVANJE KOEFICIJENATA KOD FIR SISTEMA

Analiza osetljivosti polova na promene koeficijenata imenioca, koja je izvedena u odeljku 16.3.1.1, može se direktno primeniti i na osetljivost nula funkcije prenosa na promene koeficijenata brojioca. Dakle, umesto jednačine (16.34) ima se:

$$\frac{\partial z_i}{\partial b_k} = \frac{-z_i^{M-1-k}}{\prod\limits_{\substack{j=0 \\ j \ne i}}^{M-1} (z_i - z_j)}, \quad i = 0, \ldots, M-1, \;\; k = 0, \ldots, M-1 \tag{16.56}$$

Međutim, interpretacija rezultata (16.56) sasvim je drugačija. U praksi se nule funkcije prenosa nalaze u opsegu učestanosti koji odgovara nepropusnom opsegu i nisu grupisane. Zbog toga imenilac funkcije (16.56) ne može imati male vrednosti, pa osetljivost nula na promene koeficijenata nije velika, čak ni kod direktne realizacije višeg reda. Pošto je direktna realizacija najjednostavnija, a nije mnogo osetljiva, FIR funkcije prenosa se najčešće realizuju nekom od direktnih realizacija, a vrlo retko u vidu kaskadne strukture.

Najčešći slučaj FIR funkcija prenosa u praksi su funkcije prenosa sa linearnom fazom. Kod takvih funkcija, koeficijenti zadovoljavaju uslove simetrije ili antisimetrije:

$$h[n] = \pm h[M-1-n] \tag{16.57}$$

Kako se kvantovanjem uslov (16.57) ne narušava, *uslov linearnosti faze biće zadovoljen i posle kvantovanja koeficijenata*. Naravno, amplitudska karakteristika se deformiše. Gornja granica

deformacije amplitudske karakteristike može se odrediti na sledeći način. Funkcija prenosa sa kvantovanim koeficijentima FIR sistema je:

$$\hat{H}(e^{j\Omega}) = H(e^{j\Omega}) + \Delta H(e^{j\Omega}) \tag{16.58}$$

gde je, ako se koristi aritmetika sa fiksnom tačkom i zaokružavanjem:

$$\Delta H(e^{j\Omega}) = \sum_{n=0}^{M-1} \Delta h[n] e^{-j\Omega n} \tag{16.59}$$

$$-\frac{q}{2} \le \Delta h[n] \le \frac{q}{2} \tag{16.60}$$

Iz (16.59) sledi:

$$\left|\Delta H(e^{j\Omega})\right| = \left|\sum_{n=0}^{M-1} \Delta h[n] e^{-j\Omega n}\right| \le \sum_{n=0}^{M-1} \left|\Delta h[n]\right| \left|e^{-j\Omega n}\right| \le \frac{Mq}{2} = 2^{-(B+1)} M \tag{16.61}$$

gde se za predstavljanje koeficijenata sa znakom koristi $(B+1)$ bita. Dobijena gornja granica greške amplitudske karakteristike, data sa (16.61), je suviše pesimistička jer je pretpostavljeno da sve greške koeficijenata imaju isti znak i maksimalnu vrednost. Zbog toga se tačnija potrebna dužina reči može odrediti eksperimentalno, ili statističkim pristupom, na sličan način kao kod IIR funkcija prenosa.

16.4 KVANTOVANJE PROIZVODA

Kao što je poznato, u sistemu sa fiksnom tačkom, množenjem dva binarna broja koji su predstavljeni sa $B+1$ bitom dobija se proizvod od $2B+1$ bita. Da bi se mogle vršiti dalje operacije sa dobijenim rezultatom, potrebno je skratiti rezultat na polaznu dužinu od $B+1$ bita. Skraćivanje se može izvršiti odsecanjem ili zaokruživanjem, ali se u praksi češće koristi zaokruživanje zbog toga što je srednja vrednost greške zaokruživanja jednaka nuli. Dakle, množenje digitalnog signala $x[n]$ sa koeficijentom c_i i kvantovanje rezultata može se prikazati jednačinom:

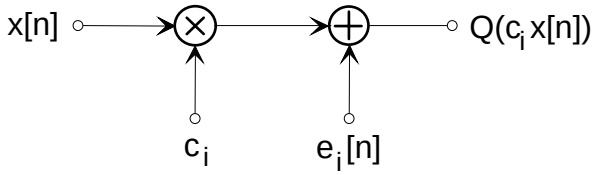
$$Q\big(c_i x[n]\big) = c_i x[n] + \varepsilon_i[n] \tag{16.62}$$

gde je $\varepsilon_i[n]$ sekvenca koja predstavlja diskretni aditivni šum koji zadovoljava sledeće uslove:

1. Greška kvantovanja $\varepsilon_i[n]$ ima uniformnu raspodelu u opsegu datom sa (16.23),
2. Greška kvantovanja $\varepsilon_i[n]$ predstavlja stacionarni beli šum,
3. Greška kvantovanja $\varepsilon_i[n]$ nije korelisana sa signalom $x[n]$,
4. Greška kvantovanja $\varepsilon_i[n]$ nije korelisana sa ulaznim signalom,
5. Greška kvantovanja $\varepsilon_i[n]$ nije korelisana sa greškom kvantovanja $\varepsilon_j[n]$, ako je $i \neq j$.

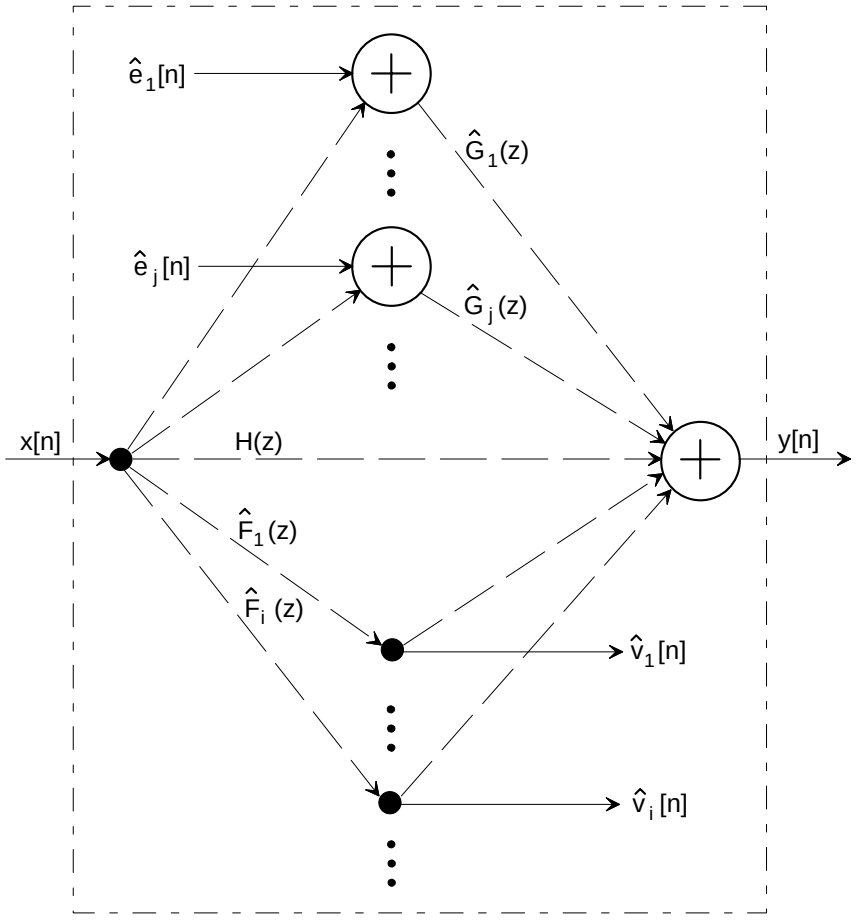
Prva tri uslova su identična sa uslovima koji su postavljeni pri analizi greške kvantovanja kod A/D konverzije. Kao i u slučaju A/D konverzije, ovi uslovi ne mogu biti zadovoljeni u svim slučajevima. Međutim, na osnovu eksperimentalnih i teorijskih analiza, ustanovljeno je da se u slučajevima složenih signala, kao što je na primer govorni signal, koji se brzo menjaju od odbirka do odbirka, navedeni uslovi mogu biti zadovoljeni. U takvim slučajevima može se primeniti linearni model šuma, kojim se jednostavno mogu odrediti osnovne statističke karakteristike šuma na izlazu kao što su srednja vrednost, varijansa, korelacione funkcije, itd.

Dakle, kvantovanje rezultata množenja se, u sistemu sa fiksnom tačkom, može prikazati modelom sa slike 16.7.



Slika 16.7 Model kvantovanja rezultata množenja.

Svakom množaču u blok dijagramu digitalnog procesora signala može se pridružiti izvor šuma $\varepsilon_i[n]$. Statističke karakteristike svih izvora šuma su iste, ali njihov doprinos ukupnom šumu na izlazu neće biti isti, jer je njihov položaj različit. Kako se u svim realizacionim strukturama, opisanim u sedmom poglavlju, izlazi množača dovode na ulaze sabirača (eventualno preko elementa za kašnjenje), ako se na sabirač dovode izlazi više množača, više izvora šuma se mogu zameniti ekvivalentnim izvorom šuma veće snage. U opštem slučaju, svi čvorovi u realizacionoj strukturi sistema za digitalnu obradu signala se mogu podeliti u dve grupe: sabirače i čvorove grananja. Izvori šuma se mogu pojaviti samo kod sabirača kao što je prikazano na slici 16.8.



Slika 16.8 Linearni model za određivanje šuma na izlazu usled kvantovanja proizvoda.

Spektralna gustina snage šuma na izlazu data je izrazom:

$$S_y(e^{j\Omega}) = \sigma_\varepsilon^2 \sum_{j=1}^{N_S} k_j \left| G_j(e^{j\Omega}) \right|^2 = \frac{q^2}{12} \sum_{j=1}^{N_S} k_j \left| G_j(e^{j\Omega}) \right|^2 \qquad (16.63)$$

gde je $q = 2^{-B}$ korak kvantovanja, k_j predstavlja broj množača čiji su izlazi vezani na ulaze *j*-tog sabirača, N_s je broj sabirača u mreži, dok je $G_j(z)$ funkcija prenosa od izlaza *j*-tog sabirača do izlaza cele mreže. Varijansa ili ukupna srednja snaga izlaznog šuma je:

$$\sigma_y^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_y(e^{j\Omega})d\Omega = \frac{q^2}{24\pi}\int_{-\pi}^{\pi}\left[\sum_{j=1}^{N_S} k_j \left|G_j(e^{j\Omega})\right|^2\right]d\Omega \tag{16.64}$$

ili, na osnovu Parsevalove teoreme,

$$\sigma_y^2 = \frac{q^2}{12}\sum_{j=1}^{N_S} k_j \sum_{n=0}^{\infty}\left|g_j[n]\right|^2 \tag{16.65}$$

Kao primer, posmatrajmo strukturu drugog reda sa polovima $z_{p1,p2} = re^{\pm j\theta}$, koja je prikazana na slici 16.4a, a čija je funkcija prenosa:

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} = \frac{1}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}} \tag{16.66}$$

odnosno, impulsni odziv:

$$h[n] = \frac{r^n}{\sin\theta}\sin(n+1)\theta\, u(n) \tag{16.67}$$

Pošto ćelija ima dva množača, postojaće dva izvora šuma, koji se sumiraju u ulaznom sabiraču sa ulaznim signalom. Dakle, u formuli (16.65) se ima $N_s = 1,\ k_1 = 2$, pa je:

$$\sigma_y^2 = \frac{q^2}{6}\sum_{n=0}^{\infty} h^2[n] = \frac{q^2}{12}\frac{1+r^2}{1-r^2}\frac{1}{r^4 - 2r^2\cos 2\theta + 1} \tag{16.68}$$

Iz izraza (16.68) se vidi da kada su polovi blizu jediničnog kruga ($r \to 1$) snaga šuma na izlazu raste, odnosno, kvari se odnos signal/šum. Situacija postaje još komplikovanija kada se primeni skaliranje funkcije prenosa, koje je potrebno da bi se izbegli nelinearni efekti. Slični rezultati se dobijaju i kada funkcija prenosa ima nule. Ono što se može u ovom trenutku zaključiti je da postoji veza između položaja polova i šuma usled kvantovanja proizvoda, što će biti predmet proučavanja u narednom odeljku.

16.4.1 SKALIRANJE KOEFICIJENATA FUNKCIJE PRENOSA

Koeficijenti funkcije prenosa, određeni postupcima sinteze iz poglavlja 8 i 9, nisu uvek pogodni za praktičnu realizaciju. Naime, u pojedinim čvorovima vrednost signala može prevazići raspoloživi dinamički opseg, koji je određen brojem raspoloživih bita, pa je potrebno smanjiti nivo signala promenom koeficijenata. Međutim, smanjenje nivoa signala kvari odnos signal/šum, pa je potrebno odrediti optimalne vrednosti koeficijenata kojima se realizuje željena funkcija prenosa kao i maksimalni odnos signal/šum.

U realizacionim strukturama digitalnih procesora signala potrebno je obezbediti da signal u svakom čvoru leži unutar dozvoljenog dinamičkog opsega. Međutim, nije potrebno vršiti kontrolu signala u svim čvorovima strukture. Ako su čvorovi spojeni granama koje predstavljaju kašnjenje, potrebno je obezbediti da ne dođe do prekoračenja opsega samo u prvom čvoru u lancu, čime se obezbeđuje zadovoljenje dinamičkog opsega u svim narednim čvorovima. Takođe, u izlaganju o

aritmetici u sistemu komplementa dvojke već je rečeno da se korektna suma dobija nezavisno od korektnosti parcijalnih suma ili sabiraka. Dakle, u ispitivanje se ne moraju uključiti čvorovi koji predstavljaju izlaze sabirača koji formiraju parcijalne sume. Čvorovi u kojima se *mora* ispitivati da li signal leži u dozvoljenom opsegu su tzv. *čvorovi grananja*, iz kojih polazi bar jedna grana koja predstavlja množač. Na slici 16.8 signali u takvim čvorovima su označeni sa $v_i[n]$, a funkcije prenosa od ulaza do takvih čvorova sa $F_i(z)$. U analizi se obično pretpostavlja da se raspoloživi dinamički opseg *M* maksimalno koristi, tj. da treba da budu zadovoljeni uslovi:

$$x[n] \le M \tag{16.69}$$

$$v_i[n] \le M, \quad i = 1,2,\ldots \tag{16.70}$$

Signali u čvorovima grananja dati su konvolucijama:

$$v_i[n] = \sum_{k=0}^{\infty} f_i[k]x[n-k], \quad i = 1,2,\ldots \tag{16.71}$$

i u principu ne moraju da zadovoljavaju uslov (16.70). Gornja granica signala $v_i[n]$ ograničena je nejednakošću:

$$\left|v_i[n]\right| \le \sum_{k=0}^{\infty} \left|f_i[k]\right|\left|x[n-k]\right| \le M\sum_{k=0}^{\infty} \left|f_i[k]\right|, \quad i = 1,2,\ldots \tag{16.72}$$

Ako se sa $\hat{v}_i[n]$ označi skalirana vrednost signala $v_i[n]$, koja takođe zadovoljava uslov (16.70), onda se iz uslova:

$$\left|\hat{v}_i[n]\right| \le M, \quad i = 1,2,\ldots \tag{16.71}$$

vidi da se skaliranje vrednosti koeficijenata originalne mreže mora izvršiti tako da bude:

$$\sum_{k=0}^{\infty} \left|\hat{f}_i[k]\right| \le 1, \quad i = 1,2,\ldots \tag{16.74}$$

gde je sa $\hat{f}_i[n]$ označen impulsni odziv, koji odgovara funkciji prenosa skaliranog kola $\hat{F}_i(z)$. Uslov (16.74) predstavlja potreban i dovoljan uslov za sprečavanje prekoračenja dinamičkog opsega. Međutim, za većinu praktičnih primena ovaj uslov je suviše strog, jer, kao što će se kasnije videti, vodi ka malom iskorišćenju dinamičkog opsega i malom odnosu signal/šum. Mnogo realniji uslovi za skaliranje koeficijenata funkcije prenosa se mogu izvesti u frekvencijskom domenu, korišćenjem tzv. L_p *normi Furijeovih transformacija*. L_p norma neke proizvoljne Furijeove transformacije $H(e^{j\omega})$ definisana je izrazom:

$$\left\|H\right\|_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|H(e^{j\Omega})\right|^p d\Omega\right]^{1/p} \tag{16.75}$$

gde je $p \ge 1$ realan broj za koji integral konvergira. Vidi se da se za $p = 1$ dobija L_1 norma, koja predstavlja srednju vrednost apsolutne vrednosti funkcije $H(e^{j\Omega})$, dok se za $p = 2$ dobija L_2 norma, koja predstavlja koren iz srednje vrednosti kvadrata funkcije $H(e^{j\Omega})$ (engl. root-mean-square - rms). Granična vrednost norme kada $p \to \infty$ predstavlja vršnu apsolutnu vrednost $H(e^{j\Omega})$, tj.

$$\left\|H\right\|_\infty = \max_{-\pi\le\Omega\le\pi} \left|H(e^{j\Omega})\right| \tag{16.76}$$

Može se pokazati da za L_p norme važi nejednakost:

$$|h[n]| \le \|H\|_1 \le \|H\|_2 \le \cdots \le \|H\|_\infty \tag{16.77}$$

Neka je ulazni signal $x[n]$ deterministički i neka njegova Furijeova transformacija $X(e^{j\omega})$ zadovoljava ograničenje:

$$\|X\|_1 \le M \tag{16.78}$$

koje zbog nejednakosti (16.77) osigurava važnost uslova (16.69). Za signale $v_i[n]$ se onda dobija:

$$v_i[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_i(e^{j\Omega})X(e^{j\Omega})e^{j\Omega n}d\Omega \tag{16.79}$$

odnosno,

$$|v_i[n]| \le \frac{1}{2\pi}\int_{-\pi}^{\pi}|F_i(e^{j\Omega})|\,|X(e^{j\Omega})|d\Omega \le \|F_i\|_\infty \frac{1}{2\pi}\int_{-\pi}^{\pi}|X(e^{j\Omega})|d\Omega \le \|F_i\|_\infty \|X\|_1 \tag{16.80}$$

Dakle, ako za ulazni signal važi ograničenje (16.78), onda se na osnovu (16.80) vidi da se prekoračenje opsega neće desiti ako se skaliranje funkcija prenosa $F_i(z)$ tako izvrši da posle skaliranja važi:

$$\|\hat{F}_i\|_\infty \le 1 \tag{16.81}$$

Skaliranje (16.81) je pogodno za sinusoidalne ulazne signale. U slučaju složenih signala, ograničenje (16.81) ne sprečava u potpunosti pojavu prekoračenja opsega kao što to čini ograničenje (16.74), ali je mnogo realnije jer je verovatnoća pojave prekoračenja mala. U primenama gde je spektar ulaznog signala širok, i ograničenje (16.81) je prestrogo, tako da se može primeniti još blaži kriterijum za ograničavanje ulaznog signala. Do ovog kriterijuma se može doći primenom nejednakosti [J-1]:

$$|v_i[n]| \le \|F_i\|_p \|X\|_q \tag{16.82}$$

gde brojevi $p, q > 1$ zadovoljavaju uslov $1/p + 1/q = 1$. Vidi se da i nejednakost (16.80) pripada ovoj klasi jer je $p = \infty$, $q = 1$. Važan slučaj od praktičnog interesa je $p = q = 2$, kada ulazni signal $x[n]$ ima konačnu energiju $E = \|X\|_2^2$, tako da se može postaviti ograničenje:

$$\|X\|_2 = \sqrt{E} \le M \tag{16.83}$$

koje dovodi do uslova za skaliranje funkcija prenosa:

$$\|\hat{F}_i\|_2 \le 1 \tag{16.84}$$

koji znači da koren iz srednje vrednosti kvadrata funkcije prenosa $\hat{F}_i(e^{j\Omega})$ mora biti manji od 1 za svako Ω. Do istog uslova se može doći i posmatranjem slučaja kada je ulazni signal $x[n]$ beli šum, a skaliranje se vrši tako da verovatnoća prekoračenja dinamičkog opsega za signal $v_i[n]$ ne bude veća od iste verovatnoće za signal $x[n]$.

Dakle, u prethodnoj analizi došlo se do tri alternativna uslova skaliranja mreže digitalnog procesora signala, kojima se izbegava prekoračenje dinamičkog opsega. Praktično izvođenje skaliranja funkcija prenosa izvodi se uvođenjem koeficijenata s_i tako da važi:

$$\hat{F}_i(z) = s_i F_i(z) \tag{16.85}$$

Za oba uslova skaliranja u frekvencijskom domenu (16.81) i (16.84), jednačina (16.85) se svodi na:

$$\left\|\hat{F}_i\right\|_p = s_i \left\|F_i\right\|_p \le 1 \tag{16.86}$$

tako da se za skalirajuće koeficijente s_i dobija:

$$s_i \le \frac{1}{\left\|F_i\right\|_p} \tag{16.87}$$

gde je $p = 2$ ili $p = \infty$. Ako je uslov skaliranja zadat u vremenskom domenu (16.74), onda je:

$$s_i \le \frac{1}{\sum_{k=0}^{\infty} |f_i[k]|} \tag{16.88}$$

U praksi se za skalirajuće koeficijente s_i obično uzimaju vrednosti $s_i' = 2^{-r} \le s_i$, čime se množenje svodi na operaciju pomeranja udesno koja je znatno brža. Ipak, u poslednje vreme se, sa pojavom digitalnih procesora signala sa integrisanim množačem, za koeficijente s_i uzimaju i tačne vrednosti date jednačinama (16.87) i (16.88), čime se, po cenu dodatnih množenja, maksimalno koristi raspoloživi dinamički opseg i popravlja odnos signal/šum.

16.4.2 ANALIZA ŠUMA KOD PARALELNE I KASKADNE REALIZACIJE

Rezultati izvedeni u prethodnom odeljku direktno se mogu primeniti na slučajeve paralelne i kaskadne realizacije digitalnih filtara, koje, kao što je naglašeno u sedmom poglavlju, predstavljaju najčešće realizacione strukture kada je red funkcije prenosa veći od dva. Pošto je paralelna realizaciona struktura nešto jednostavnija za analizu šuma, prvo će biti razmotren slučaj paralelne realizacije.

16.4.2.1 Analiza šuma kod paralelne realizacije

Kao što je bilo rečeno u odeljku 7.3.4, paralelna realizacija se sastoji od paralelne veze ćelija drugog reda koje su prikazane na slici 7.17. Posmatrajmo prvo slučaj kada se koristi direktna kanonička ćelija sa slike 7.17b, koji je ujedno i najčešći u praksi. U takvoj realizaciji (označimo je sa P1) potrebno je skalirati signal u samo jednom čvoru u svakoj ćeliji, kao i signal u izlaznom čvoru. Skaliranje izlaznog signala se može izbeći skaliranjem cele funkcije prenosa tokom projektovanja. Paralelna realizacija koja koristi takve ćelije prikazana je na slici 16.9 na kojoj su označeni izvori šuma kao i parcijalne funkcije prenosa $\hat{F}_i(z)$ i $\hat{G}_i(z)$ i skalirani signali $\hat{v}_i[n]$.

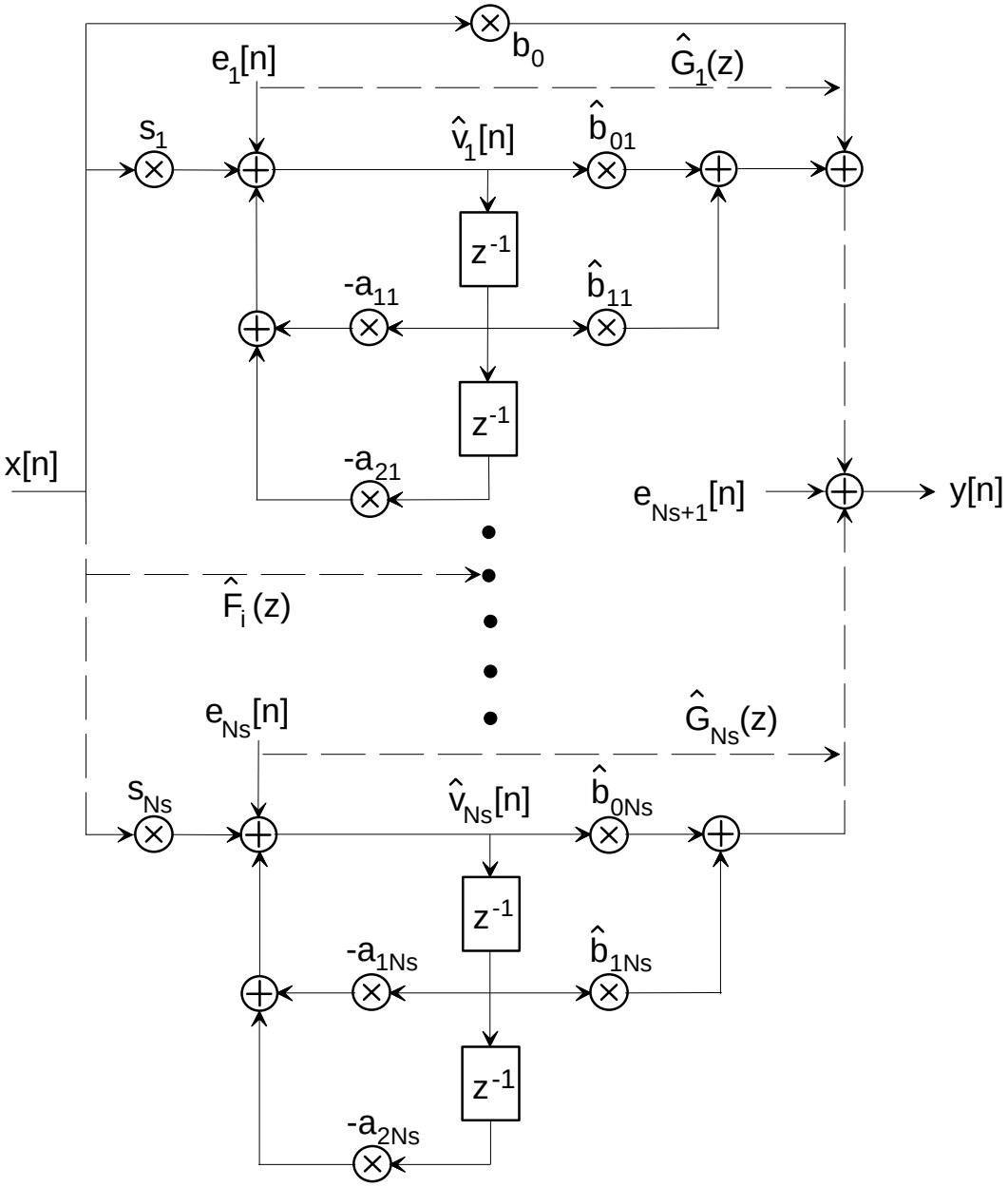
Skaliranje signala $v_i[n]$ se najjednostavnije izvodi uvođenjem množača s_i na ulazu u svaku ćeliju i korekcijom vrednosti množača b_{0i} i b_{1i} da se funkcija prenosa ćelije ne bi promenila. Dakle, imamo:

$$\hat{b}_{0i} = \frac{b_{0i}}{s_i}, \quad \hat{b}_{1i} = \frac{b_{1i}}{s_i}, \quad i = 1, \ldots, N_s \tag{16.89}$$

a za skalirane funkcije prenosa $\hat{G}_i(z)$ se dobija:

$$\hat{G}_i(z) = \frac{\hat{b}_{0i} + \hat{b}_{1i} z^{-1}}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}}, \quad i = 1, \ldots, N_s \tag{16.90}$$

dok je $\hat{G}_{N_s+1}(z) = 1$.



Slika 16.9 Skaliranje paralelne realizacije P1 sa direktnim kanoničkim ćelijama.

Funkcije prenosa pre skaliranja $F_i(z)$ date su izrazima:

$$F_i(z) = \frac{1}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}}, \quad i = 1, \ldots, N_s \tag{16.91}$$

a sa slike 16.9 se vidi da je posle skaliranja $\hat{F}_i(z) = s_i F_i(z)$, čime se smanjuje veliko pojačanje $\|F_i\|_p$.

Sa slike 16.9 se takođe vidi da u svakoj sekciji postoji samo jedan izvor šuma kao i jedan izvor šuma u izlaznom sabiraču. Ako se zaokruživanje ili odsecanje rezultata množenja vrši pre sabiranja, onda je u formulama (16.63) i (16.64):

$$\begin{aligned} k_i &= 3, \quad i = 1, \ldots, N_s - 1 \\ k_{N_s} &= 3 - (2N_s - N) \\ k_{N_s+1} &= N + 1 \end{aligned} \tag{16.92}$$

U slučaju da se zaokruživanje ili odsecanje vrši tek posle sabiranja rezultata množenja, onda je $k_i = 1$ za svako i.

Ako su koeficijenti skaliranja pojedinih ćelija tako izabrani da relacija (16.87) predstavlja jednakost, onda iz (16.89) sledi:

$$\hat{G}_i(z) = \frac{1}{s_i} G_i(z) = \|F_i\|_p \, G_i(z), \quad i = 1, \ldots, N_s \tag{16.93}$$

tako da je, prema (16.63), gustina spektra snage šuma data izrazom:

$$S_y(e^{j\Omega}) = \sigma_\varepsilon^2 \left[k_{N_s+1} + \sum_{j=1}^{N_S} k_j \|F_j\|_p^2 \left| G_j(e^{j\Omega}) \right|^2 \right] \tag{16.94}$$

Iz izraza (16.94) se vidi veza između skaliranja funkcije prenosa i šuma. Ako su polovi blizu jediničnog kruga, faktor $\|F_i\|_p$ biće veliki, tako da se šum povećava. Zbog toga se i izbegava skaliranje na osnovu suviše pesimističke norme u vremenskom domenu (11.88), a takođe se u (16.87) radije bira $p = 2$ umesto $p = \infty$, ako je to moguće. Iz (16.94) se može dobiti i srednja snaga šuma na izlazu:

$$\sigma_y^2 = \sigma_\varepsilon^2 \left[k_{N_s+1} + \sum_{j=1}^{N_S} k_j \|F_j\|_p^2 \|G_j\|_2^2 \right] \tag{16.95}$$

Drugi oblik paralelne realizacione strukture (P2) koristi ćelije drugog reda koje su prikazane na slici 7.17d, a koje su dobijene transpozicijom ćelije sa slike 7.17b, prikazan je na slici 16.10. Skaliranje se obavlja promenom vrednosti koeficijenata množača b_{0i} i b_{1i}, a kompenzacija skaliranja uvođenjem dodatnih množača iza ćelija drugog reda. Dakle, imamo:

$$\hat{b}_{0i} = s_i b_{0i}, \quad \hat{b}_{1i} = s_i b_{1i}, \quad i = 1, \ldots, N_s \tag{16.96}$$

Analiza skaliranja izvedena za paralelnu realizaciju P1 i dalje važi, ali je sada zbog transpozicije ćelije:

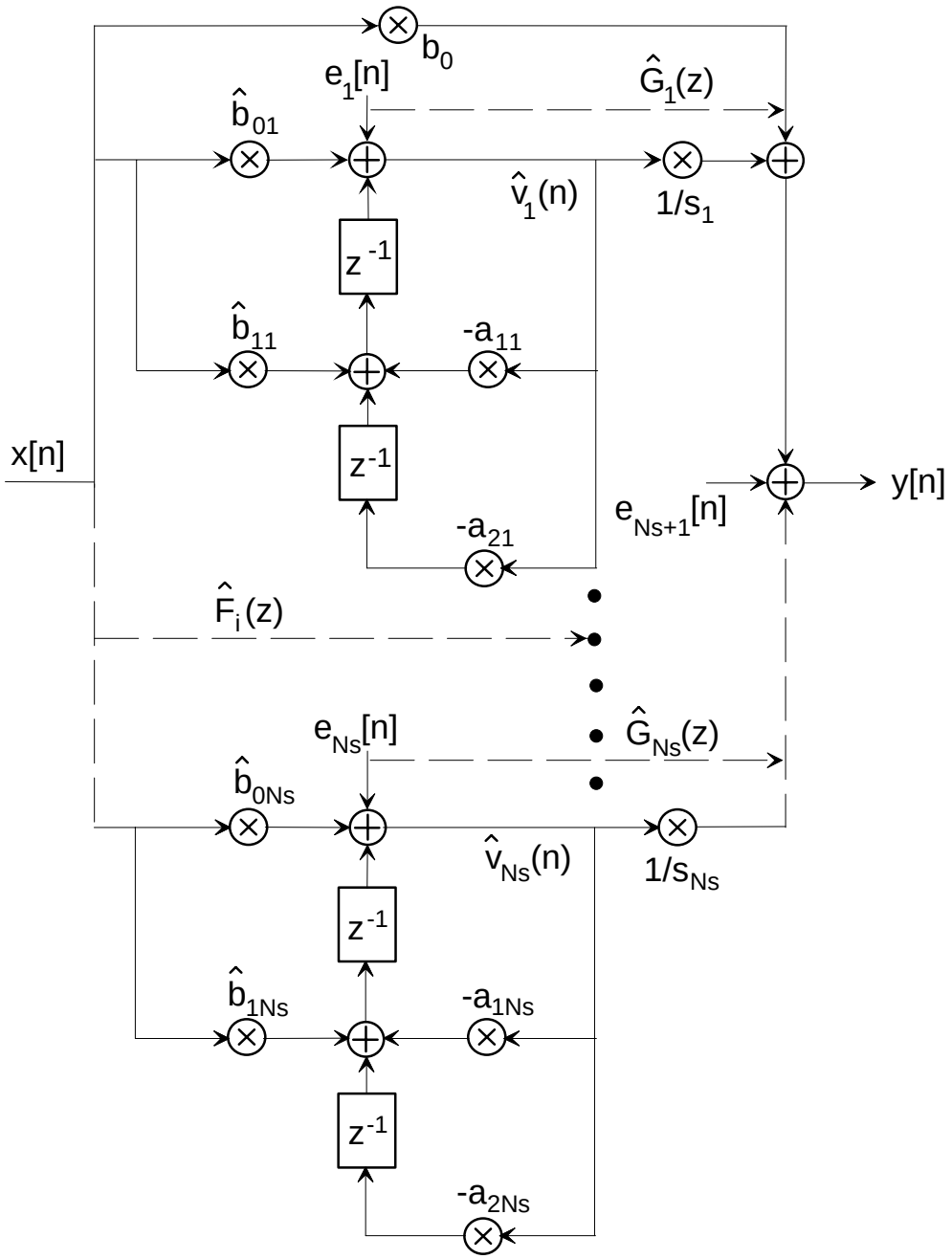
$$F_i(z) = \frac{\hat{b}_{0i} + \hat{b}_{1i} z^{-1}}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}}, \quad i = 1, \ldots, N_s \tag{16.97}$$

$$\hat{G}_i(z) = \frac{1}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}}, \quad i = 1, \ldots, N_s \tag{16.98}$$

Izrazi za gustinu snage i varijansu šuma na izlazu (16.94) i (16.95) i dalje važe, s tim što su vrednosti koeficijenata k_i kada se zaokruživanje obavlja pre sumiranja dati izrazom:

$$\begin{aligned} k_i &= 4, \quad i = 1, \ldots, N_s - 1 \\ k_{N_s} &= 4 - 2(2N_s - N) \\ k_{N_s+1} &= N_s + 1 \end{aligned} \tag{16.99}$$

Ako se zaokruživanje vrši posle množenja sa $1/s_i$ i sumiranja, onda je $k_i = 1$ za svako i.



Slika 16.10 Skaliranje paralelne realizacije P2 sa transponovanim kanoničkim ćelijama.

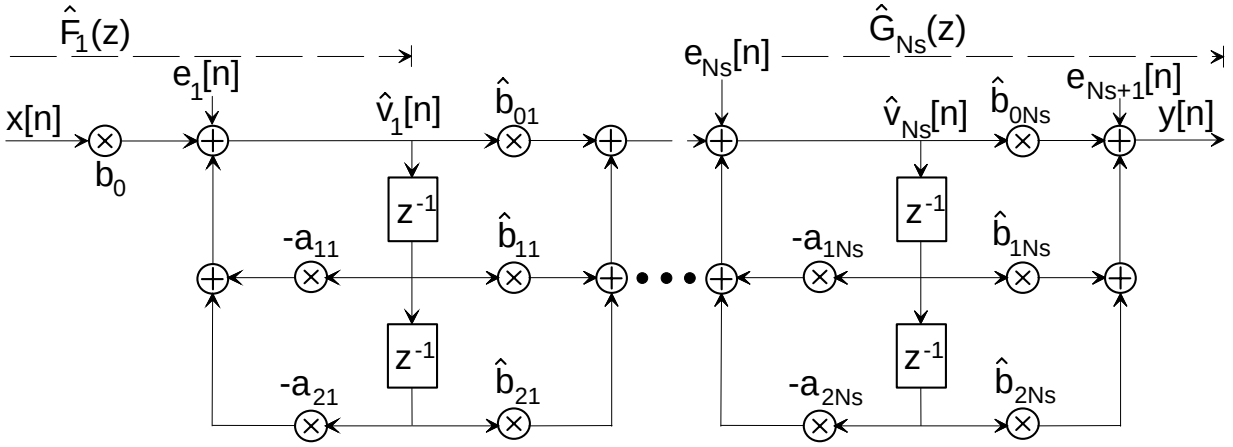
Kao što se vidi iz (16.95), za $p = 2$ jedina razlika između dve paralelne realizacije potiče od različitih vrednosti koeficijenata k_i. Snaga šuma je veća kod P2 realizacije za $4/3$ puta (1.25 dB). Ako je $p \neq 2$, pojavljuju se dodatne razlike, koje potiču od različitih definicija parcijalnih funkcija prenosa $\hat{F}_i(z)$ i $\hat{G}_i(z)$. Ipak, ove razlike nisu tako značajne, tako da je i u tim slučajevima P1 realizacija bolja za oko 1.25 dB u pogledu snage generisanog šuma. Dakle, *ako nema dodatnih zahteva, treba koristiti paralelnu realizaciju P1 jer je bolja u pogledu generisanog šuma*.

16.4.2.2 Analiza šuma kod kaskadne realizacije

U slučaju kaskadne realizacije, analiza šuma i izbor strukture su znatno komplikovaniji. Naime, pored analize generisanog šuma usled kvantovanja proizvoda i uticaja skaliranja, kod kaskadne realizacije treba odrediti i optimalno sparivanje faktora drugog reda u imeniocu i brojiocu parcijalnih funkcija prenosa kao i optimalan redosled sekcija drugog reda. U ovom odeljku će biti

izvedena samo analiza šuma, dok će problem optimalne dekompozicije i redosleda sekcija biti razmatran u sledećem odeljku.

Jedna od najčešće korišćenih verzija kaskadne strukture je direktna kanonička kaskadna struktura (K1), u kojoj se koriste ćelije sa slike 7.15b. Skaliranje se ostvaruje promenom koeficijenata množača b_{0i}, b_{1i} i b_{2i} i uvođenjem dodatnog ulaznog množača b_0, kao što je prikazano na slici 16.11.



Slika 16.11 Skaliranje kaskadne realizacije K1 sa direktnim kanoničkim ćelijama.

Za razliku od paralelnih realizacija P1 i P2, kod kaskadne realizacije funkcije prenosa $\hat{F}_i(z)$ i $\hat{G}_i(z)$ i-te sekcije sadrže funkcije prenosa prethodnih i narednih sekcija što otežava analizu. Sa slike 16.11 se vidi da je:

$$\hat{F}_i(z) = \frac{\hat{b}_0}{1 + a_{1i}z^{-1} + a_{2i}z^{-2}} \prod_{k=1}^{i-1} \hat{H}_k(z), \quad i = 1, \ldots, N_s \tag{16.100}$$

$$\hat{G}_i(z) = \prod_{k=i}^{N_s} \hat{H}_k(z), \quad i = 1, \ldots, N_s \tag{16.101}$$

gde je sa $\hat{H}_k(z)$ označena funkcija prenosa k-te sekcije. U svakoj sekciji postoji samo jedan izvor šuma, kao i izvor šuma na izlazu. Ako se koeficijenti skaliranja izaberu tako da relacija (16.87) predstavlja jednakost, onda su skalirani koeficijenti množača u kaskadnoj realizaciji K1 dati izrazima:

$$\begin{aligned} \hat{b}_0 &= s_i \\ \hat{b}_{ki} &= \frac{s_{i+1}}{s_i} b_{ki}, \quad k = 0,1,2, \quad i = 1, \ldots, N_s \end{aligned} \tag{16.102}$$

gde je $s_{N_s+1} = b_0$ da bi se ostvarilo željeno pojačanje. Izraz za parcijalnu funkciju prenosa $F_i(z)$ pre skaliranja je:

$$F_i(z) = \frac{1}{1 + a_{1i}z^{-1} + a_{2i}z^{-2}} \prod_{k=1}^{i-1} H_k(z), \quad i = 1, \ldots, N_s \tag{16.103}$$

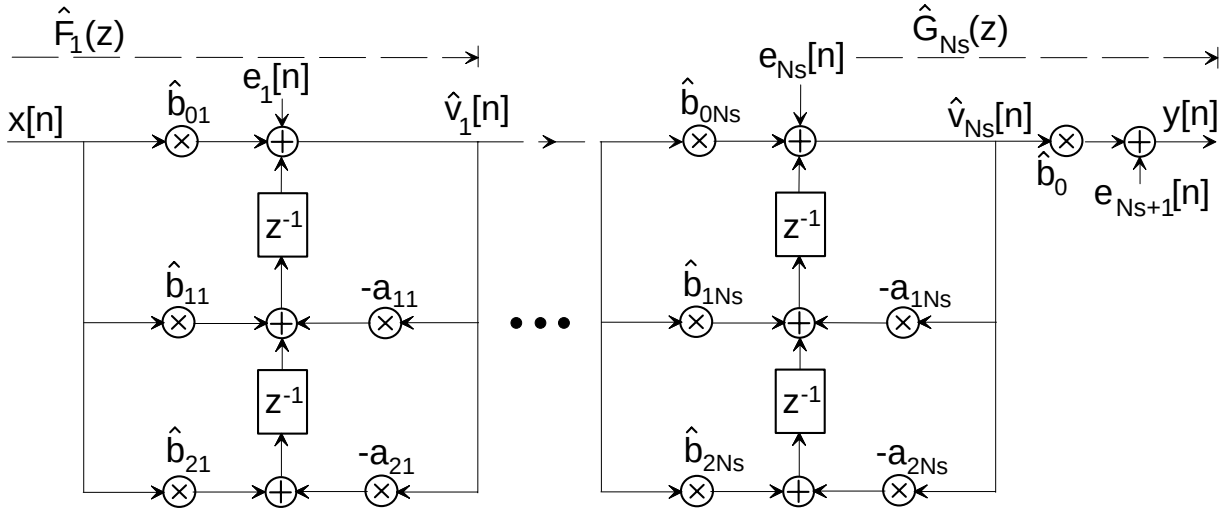
Lako se može pokazati da su gustina snage i varijansa šuma na izlazu dati izrazima (16.94) i (16.95), gde je:

$$\begin{aligned} k_1 &= 3, \\ k_i &= 5 \quad i = 2,\ldots,N_s-1 \\ k_{N_s} &= 5-(2N_s-N) \\ k_{N_s+1} &= 3-(2N_s-N) \end{aligned} \tag{16.104}$$

Na kraju, razmotrimo slučaj alternativne kaskadne realizacije (K2) sa slike 16.12, koja se dobija ako se upotrebe ćelije sa slike 7.15d. Kao što je poznato, ova ćelija se dobija transpozicijom ćelije sa slike 7.15b. Međutim, realizacija sa slike 16.12 ne predstavlja pravu transpoziciju strukture sa slike 16.11, jer redosled sekcija nije promenjen. Zbog toga je:

$$\hat{F}_i(z) = \prod_{k=1}^{i} \hat{H}_k(z), \quad i = 1,\ldots,N_s \tag{16.105}$$

$$\hat{G}_i(z) = \frac{\hat{b}_0}{1+a_{1i}z^{-1}+a_{2i}z^{-2}} \prod_{k=i+1}^{N_s} \hat{H}_k(z), \quad i = 1,\ldots,N_s \tag{16.106}$$



Slika 16.12 Skaliranje kaskadne realizacije K2 sa transponovanim kanoničkim ćelijama.

Skaliranje koeficijenata se obavlja prema formulama:

$$\begin{aligned} \hat{b}_{ki} &= \frac{s_i}{s_{i-1}} b_{ki}, \quad k = 0,1,2, \quad i = 1,\ldots,N_s \\ \hat{b}_0 &= \frac{b_0}{s_{N_s}} \end{aligned} \tag{16.107}$$

gde je $s_0 = 1$.

Izrazi za izlazni šum (16.94) i (16.95) i dalje važe, ali sa težinskim koeficijentima:

$$\begin{aligned} k_i &= 5, \quad i = 1,\ldots,N_s-1 \\ k_{N_s} &= 5-2(2N_s-N) \\ k_{N_s+1} &= 1 \end{aligned} \tag{16.106}$$

Iz prethodne analize mogu se izvesti zaključci koju od kaskadnih realizacija K1 i K2 treba upotrebiti u nekom određenom slučaju. Odgovor zavisi od toga koja se norma spektralne gustine šuma $\|S_y\|_r$ minimizira kao i od izabranog kriterijuma skaliranja $c_i \le 1/\|F_i\|_p$. Iz izraza (16.94) i

(16.106) se vidi da se u pogledu šuma realizacije K1 i K2 razlikuju samo po tome da li je dodatni faktor u imeniocu uključen u $\hat{F}_i(z)$ u slučaju K1 realizacije, ili u $\hat{G}_i(z)$ u slučaju K2 realizacije. Ispitivanja sprovedena u [J–2] pokazuju da su u slučajevima $p=2,\ r=1$ i $p=\infty,\ r=\infty$ obe konfiguracije praktično identične u pogledu šuma, da je u slučaju $p=2,\ r=\infty$ realizacija K1 bolja, a da je u slučaju $p=\infty,\ r=1$ realizacija K2 bolja.

U literaturi [J-4] izveden je još jedan interesantan rezultat koji se odnosi na minimalnu vrednost izlaznog šuma. Ako se posmatra bilo koja od četiri analizirane konfiguracije P1, P2, K1 i K2, vidi se da množači u kolu reakcije a_{ki} i-te sekcije imaju kao ulazni signal $v_i[n]$ ili njegovu zakašnjenu verziju. Šum koji se dobija kvantovanjem rezultata množenja predstavlja jednu komponetnu izvora šuma $\varepsilon_i[n]$. Funkcija prenosa od ulaza kola do ulaza množača je $z^{-q}F_i(z)$, a od izlaza množača do izlaza kola je $z^{-r}G_i(z)$, gde je $q+r=k$. Ako se koeficijent a_{ki} malo promeni za Δa_{ki}, funkcije prenosa $F_i(z)$ i $G_i(z)$ se praktično ne menjaju, dok se ukupna funkcija prenosa $H(z)$ približno menja za iznos:

$$\Delta H(z) \approx z^{-k}F_i(z)G_i(z)\Delta a_{ki} \tag{16.109}$$

Iz izraza (16.109) se dobija osetljivost funkcije prenosa $H(z)$ na promene koeficijenta množača a_{ki} kao:

$$\frac{\partial H(z)}{\partial a_{ki}} = z^{-k}F_i(z)G_i(z) \tag{16.110}$$

Pošto član z^{-k} unosi čisto kašnjenje i ne utiče na bilo koju L_p normu osetljivosti, pogodnije je definisati funkciju osetljivosti bez kašnjenja kao:

$$S_i(z) = F_i(z)G_i(z) \tag{16.111}$$

Na osnovu Švarcove nejednakosti lako je pokazati da važi:

$$\|S_i\|_1 = \|F_iG_i\|_1 \le \|F_i\|_2 \cdot \|G_i\|_2 \tag{16.112}$$

odnosno, na osnovu (16.95), dobija se donja granica za šum kada se skaliranje vrši na osnovu norme L_2:

$$\sigma_y^2 \ge \sigma_\varepsilon^2\left[k_{N_s+1} + \sum_{i=1}^{N_S} k_i \|S_i\|_1^2\right],\ \text{za}\ \ p=2 \tag{16.113}$$

Takođe se iz (16.111) dobija:

$$\|S_i\|_2 = \|F_iG_i\|_2 \le \|F_i\|_\infty \cdot \|G_i\|_2 \tag{16.114}$$

odakle se dobija donja granica za šum kada se skaliranje vrši na osnovu norme L_∞:

$$\sigma_y^2 \ge \sigma_\varepsilon^2\left[k_{N_s+1} + \sum_{i=1}^{N_S} k_i \|S_i\|_2^2\right],\ \text{za}\ \ p=\infty \tag{16.115}$$

Takođe se iz relacije:

$$\left|S_i(e^{j\Omega})\right| = \left|F_i(e^{j\Omega})G_i(e^{j\Omega})\right| \le \|F_i\|_\infty \cdot \left|G_i(e^{j\Omega})\right| \tag{16.116}$$

dobija donja granica za spektralnu gustinu snage izlaznog šuma kada se skaliranje vrši na osnovu norme L_∞:

$$S_y(e^{j\Omega}) \geq \sigma_\varepsilon^2 \left[k_{N_s+1} + \sum_{i=1}^{N_S} k_i \left| S_i(e^{j\Omega}) \right| \right], \quad \text{za } p = \infty \tag{16.117}$$

Tačnost formula za granične vrednosti (16.113), (16.115) i (16.117) je dobra, jer je greška manja od 6 dB. Izvedene granične vrednosti malo zavise od sparivanja nula i polova i redosleda sekcija u kaskadnoj realizaciji, kao i od transponovanja strukture ćelija, jer ove operacije utiču na rezultat samo preko koeficijenata k_i. Zbog toga se granične vrednosti mogu izračunati pre skaliranja kola i tako izvršiti procena da li nivo izlaznog šuma zadovoljava postavljene kriterijume.

16.4.3 OPTIMIZACIJA KASKADNE REALIZACIJE

U prethodnom odeljku uočeno je, da kod kaskadne realizacije, snaga generisanog šuma usled kvantovanja proizvoda zavisi i od sparivanja faktora drugog reda u imeniocu i brojiocu parcijalnih funkcija prenosa kao i od redosleda sekcija drugog reda. Određivanje optimalnog sparivanja nula i polova i optimalnog redosleda je vrlo složen problem, koji je dugo rešavan u teoriji aktivnih filtara, ali sa malo uspeha. Naime, lako se pokazuje da postoji $(N_s)!$ načina sparivanja kao i $(N_s)!$ mogućih redosleda sekcija. Ako broj sekcija N_s nije mali, ispitivanje $(N_s!)^2$ mogućih realizacija je ogroman posao, koji se teško može obaviti čak i uz pomoć računara. U literaturi o aktivnim filtrima je predloženo više načina za jednostavnije rešenje problema optimizacije, ali je njihova zajednička odlika da su suviše komplikovani za praktičnu primenu. Zbog toga su razvijeni heuristički metodi, kojima se nezavisno rešavaju problemi optimizacije sparivanja i redosleda, a koji se sa malim modifikacijama mogu primeniti i u realizaciji digitalnih funkcija prenosa.

Razmotrimo prvo problem sparivanja nula i polova funkcije prenosa. Posmatrajući izraze za generisani šum (16.94) i (16.95), kao i izraze za parcijalne funkcije prenosa $F_i(z)$ i $G_i(z)$ pre skaliranja, uočava se da se funkcija prenosa jedne sekcije $H_i(z)$ pojavljuje bilo u $F_i(z)$ bilo u $G_i(z)$ ali ne u obe. Dakle, ako se želi minimizacija šuma, potrebno je minimizirati neku normu funkcije prenosa $H_i(z)$. Iz opštih osobina funkcije prenosa digitalnog sistema takođe sledi da se polovi funkcije prenosa u z-ravni moraju nalaziti u blizini propusnog opsega i prema tome unose pojačanje. Najveće pojačanje unose polovi koji su najbliži jediničnom krugu (ekvivalentno polovima sa najvećim Q faktorom kod analognih filtara). S druge strane, nule funkcije prenosa unose slabljenje i moraju se nalaziti dalje od propusnog opsega. Na osnovu ovoga se može zaključiti da se *na izlazu dobija najmanji šum ako je amplitudska karakteristika sekcije što ravnija u propusnom opsegu, što se dobija sparivanjem nula i polova koji su najbliži*. Dakle, optimalno sparivanje se može dobiti na sledeći način. Prvo se uoči par polova najbliži jediničnom krugu i par nula koji mu je najbliži i formira se funkcija prenosa $H_1(z)$. Zatim se uoči sledeći par polova najbliži jediničnom krugu i spari sa najbližim parom nula. Postupak se nastavlja sve dok se ne izvrši sparivanje svih polova i nula.

Ovaj jednostavan kriterijum sparivanja se u praksi pokazao kao veoma dobar i u većini slučajeva predstavlja optimalno rešenje.

Razmatranje problema redosleda sekcija još je komplikovanije. Problem se svodi na izbor da li će se funkcija prenosa neke sekcije $H_i(z)$ pojavljivati češće u $F_i(z)$ ili u $G_i(z)$, odnosno, da li će biti postavljena bliže ulazu ili izlazu. Rešenje zavisi i od toga koje se norme primenjuju na $F_i(z)$ i $G_i(z)$. Ako je $p = 2$, obe norme su tipa L_2, pa redosled može biti proizvoljan. U opštem slučaju,

rezultat zavisi od izabranih vrednosti za p i r koje definiše normu $||S_y||_r$. Pogodno je definisati meru varijacije pojačanja sekcije na sledeći način:

$$V_i = \frac{\|H_i\|_\infty}{\|H_i\|_2} \tag{16.118}$$

Ako je $p=\infty,\ r=1$ *pokazuje se da je optimalan redosled takav da su sekcije sa najvećom varijacijom pojačanja* (*sa polovima najbližim jediničnom krugu*) *postavljene na kraju kaskade*, tj. da se češće pojavljuju u $G_i(z)$ nego u $F_i(z)$. Međutim, može se pokazati, da *u slučaju* $p=2,\ r=\infty$, tj. kada se minimizira norma spektralne gustine šuma $||S_y||_\infty$, *važi upravo obrnuti kriterijum da su sekcije sa najvećom varijacijom pojačanja postavljene na početku kaskade*. *U ostala dva slučaja od interesa* $p=2,\ r=1$ *i* $p=\infty,\ r=\infty$ *redosled zavisi i od toga koja je kaskadna realizacija upotrebljena*. *Ako se koristi realizacija K2, optimalno je da sekcije sa najvećom varijacijom pojačanja budu postavljene na početku kaskade, a ako se koristi realizacija K1, na kraju kaskade*.

Sada se može izvršiti i kompletnije poređenje paralelne i kaskadne realizacije koje uključuje i analizu šuma. Paralelnu strukturu je jednostavnije projektovati jer se ne vrši sparivanje nula i polova niti određivanje redosleda, a i postupak skaliranja je jednostavniji. Osim toga, eksperimentalni rezultati pokazuju da je nivo šuma na izlazu paralelne realizacije P1 istog reda kao kod optimalnih kaskadnih realizacija. Uticaj kvantovanja koeficijenata takođe je manji kod paralelne realizacije [K-16]. Dakle, moglo bi se reći da je paralelna realizacija P1 optimalna realizaciona struktura. Međutim, i kaskadna realizacija ima svojih prednosti. Pre svega, u najčešćem slučaju kada se sinteza vrši bilinearnom transformacijom klasičnih analognih funkcija prenosa, nule digitalne funkcije prenosa leže na jediničnom krugu pa se može uštedeti 25% ili čak i 50% množača. Drugo, nule se ne pomeraju sa jediničnog kruga zbog kvantovanja koeficijenata, što povoljno utiče na amplitudsku karakteristiku u nepropusnom opsegu.

16.4.4 ANALIZA ŠUMA KOD FIR FUNKCIJA PRENOSA

Analiza šuma usled kvantovanja proizvoda kod realizacionih struktura FIR funkcija prenosa najčešće je znatno jednostavnija nego kod IIR funkcija prenosa. Naime, opšti izraz za spektralnu gustinu snage izlaznog šuma usled kvantovanja proizvoda (16.63) važi i dalje. Kako se FIR sistemi najčešće realizuju direktnom i transponovanom direktnom realizacijom, svi izvori šuma se mogu pridružiti izlaznom sabiraču, tako da je u formulama (16.63) i (16.65) $k_j=1$, $|G_j(e^{j\Omega})|=1$ i $N_s=M$, pa je:

$$\sigma_y^2 = S_y(e^{j\Omega}) = M\frac{q^2}{12} = M\frac{2^{-2B}}{12} \tag{16.119}$$

ako je dužina reči $B+1$ bita, a kvantovanje proizvoda se vrši zaokružavanjem pre sabiranja. Ako se proces zaokružavanja obavlja posle sabiranja, onda je:

$$\sigma_y^2 = S_y(e^{j\Omega}) = \frac{q^2}{12} = \frac{2^{-2B}}{12} \tag{16.120}$$

Ako FIR sistem koji se realizuje ima linearnu fazu, brojevi množača i izvora šuma u realizacionoj strukturi su oko dva puta manji, pa se u (16.119) M zamenjuje sa $\lceil (M+1)/2 \rceil$.

Što se tiče skaliranja koeficijenata radi sprečavanja prekoračenja opsega, ako se koristi predstavljanje brojeva u sistemu komplementa dvojke, potrebno je voditi računa samo o veličini izlaznog signala, jer svi ostali signali predstavljaju parcijalne sume.

16.4.5 ANALIZA ŠUMA KOD ARITMETIKE SA POKRETNOM TAČKOM

Iz prethodnih izlaganja se vidi da je glavni nedostatak realizacije sistema za digitalnu obradu signala, koji koristi reprezentaciju brojeva sa fiksnom tačkom, ograničen dinamički opseg. To za sobom povlači komplikovane postupke skaliranja koeficijenata, što negativno utiče na šum koji se generiše u sistemu. Većina ovih problema može se rešiti korišćenjem reprezentacije brojeva sa pokretnom tačkom.

Analiza šuma kod sistema za digitalnu obradu signala koji koriste aritmetiku sa pokretnom tačkom znatno je složenija nego ista analiza za aritmetiku sa fiksnom tačkom. Razlog za to je činjenica da je greška kvantovanja broja predstavljenog sa pokretnom tačkom srazmerna vrednosti samog broja, tj. generisani šum je multiplikativne prirode. Zbog toga pretpostavke da su izvori šuma statistički nezavisni od signala i da generišu beli šum nisu više opravdane. Osim toga, zbog toga što se normalizacija (kvantovanje) mantise vrši i posle množenja i posle sabiranja, *u kolo se moraju ubaciti izvori šuma posle svakog množenja i posle svakog sabiranja*. Redosled izvođenja operacija množenja i sabiranja može u velikoj meri uticati na generisanje šuma.

U analizi ponašanja sistema koji koriste aritmetiku sa pokretnom tačkom mora se napraviti pretpostavka o prirodi ulaznog signala. Obično se kao ulazni signal koristi beli šum, jer se tada opravdano može pretpostaviti da je greška kvantovanja statistički nezavisna od ulaznog signala.

U literaturi [S-2], [L-11], [W-4], [K-5] izvršena je detaljna analiza grešaka kvantovanja u sistemu sa pokretnom tačkom i dobijeni mnogi korisni rezultati. Najvažniji rezultati se odnose na poređenje sistema sa fiksnom tačkom i sistema sa pokretnom tačkom, koji za mantisu koristi isti broj bita kao sistem sa fiksnom tačkom. Pokazano je da pod tim uslovima aritmetika sa pokretnom tačkom daje bolji odnos signal-šum na izlazu, naročito u slučajevima kada se polovi sistema nalaze blizu jediničnog kruga. Naravno, cena koja je plaćena leži u većoj kompleksnosti sistema zbog dodatnih bitova za predstavljanje eksponenta, kao i zbog složenijeg hardvera za obavljanje aritmetičkih operacija sa brojevima sa pokretnom tačkom. Pored boljih karakteristika u pogledu šuma, aritmetika sa pokretnom tačkom potpuno eliminiše problem prekoračenja opsega, što omogućava jednostavnije projektovanje sistema za obradu signala.

16.5 NELINEARNI EFEKTI

Ako se u nekom trenutku $n = n_0$ ukine pobudni signal stabilnog IIR diskretnog sistema, izlazni signal bi trebalo da asimptotski opada ka nuli. Međutim, ako se u realizaciji IIR sistema za predstavljanje signala i koeficijenata koristi konačan broj bita, izlazni signal može oscilovati ili imati konstantnu vrednost različitu od nule. Taj efekat se naziva *granični ciklus pri nultoj pobudi* i posledica je nelinearnih pojava kod kvantovanja proizvoda ili prekoračenja opsega kod sabiranja.

Nelinearne pojave je u opštem slučaju vrlo teško analizirati, pa se pribegava korišćenju linearnih modela nelinearnih pojava ili se egzaktno analiziraju vrlo prosti sistemi. Zbog toga će analiza nelinearnih efekata u ovom odeljku biti ograničena samo na sisteme prvog i drugog reda.

Ipak, izvedeni rezultati se mogu generalizovati, a takođe su praktično upotrebljivi kada se sistem realizuje u vidu paralelne ili kaskadne strukture.

16.5.1 GRANIČNI CIKLUSI ZBOG KVANTOVANJA PROIZVODA

U analizi procesa kvantovanja u odeljku 16.4 bilo je pretpostavljeno da se greška kvantovanja može predstaviti diskretnom sekvencom, koja zadovoljava pet uslova. U tom slučaju greška kvantovanja predstavlja aditivni beli šum. Međutim, ako je nivo signala u kolu mali i predstavljen je samo sa nekoliko susednih kvantizacionih nivoa (a to se dešava kada se ukine pobuda), onda ne važe pretpostavke 3 i 5 koje se odnose na korelaciju grešaka. Naime, kada vrednosti signala uzimaju nekoliko susednih vrednosti, korelisanost grešaka istog množača u raznim trenucima, ili korelisanost grešaka različitih množača, se sigurno povećava. U takvim slučajevima sistem počinje da se ponaša nelinearno.

Posmatrajmo, kao primer, sistem prvog reda opisan diferencnom jednačinom:

$$y[n] = ay[n-1] + x[n], \quad |a| < 1 \tag{16.121}$$

koji treba, radi jednostavnije analize, realizovati sa četvorobitnom (3+1) aritmetikom. Pretpostavimo da se proizvod $ay[n-1]$ kvantuje zaokružavanjem pre sabiranja sa signalom $x[n]$. Onda se diferencna jednačina (16.121) može napisati u obliku:

$$\hat{y}[n] = Q\left(a\hat{y}[n-1]\right) + \hat{x}[n] \tag{16.122}$$

Neka je pobudni signal $\hat{x}[n] = 0.875_{10}\delta[n] = 0.111_2\delta[n]$, a koeficijent množača (pol sistema) $a = 0.5_{10} = 0.100_2$. Zamenom vrednosti $n = 0,1,2,\ldots$ iz jednačina (16.121) i (16.122) se mogu dobiti izlazni signali iz idealnog sistema bez kvantovanja proizvoda i realnog sistema sa kvantovanjem proizvoda svođenjem na 4 bita. Rezultati su prikazani u drugoj i trećoj koloni Tabele 16.2.

Tabela 16.2 Rezultati diferencnih jednačina (16.121) i (16.122).

n	$y[n]$	$\hat{y}[n]$	$y[n]$	$\hat{y}[n]$
0	0.875	0.875	0.875	0.875
1	0.4375	0.500	-0.4375	-0.500
2	0.21875	0.250	0.21875	0.250
3	0.109375	0.125	-0.109375	-0.125
4	0.0546875	0.125	0.0546875	0.125
5	0.02734375	0.125	-0.02734375	-0.125

Kao što se vidi, izlazni signal iz sistema sa kvantovanjem proizvoda se zaustavlja na konačnoj vrednosti $\hat{y}[n] = 0.125$, za $n \geq 3$. Takva pojava naziva se *granični ciklus*.

Ako je u sistemu opisanom sa (16.121) i (16.122) koeficijent množača $a = -0.5$, dobijaju se rezultati prikazani u četvrtoj i petoj koloni Tabele 16.2. U ovom slučaju sistem osciluje, a izlazni signal uzima samo dve vrednosti 0.125 i -0.125.

Lako se može uočiti da je u oba slučaja IIR sistem, koji je bio stabilan kada nema kvantovanja proizvoda, postao nestabilan kada se proizvod kvantuje. Vrednosti signala koje takav sistem može imati u graničnom ciklusu nazivaju se *mrtve zone*.

Uticaj kvantovanja se može protumačiti i na drugačiji način. Naime, nestabilnost sistema odgovara pomeranju pola sistema na jedinični krug. U ova dva slučaja, pol sistema koji se nalazio u

tački $a = \pm 0.5$, se pomerio u tačku $a = \pm 1$, čime je sistem postao nestabilan. Koristeći ovu interpretaciju graničnog ciklusa mogu se dobiti generalni rezultati za sistema prvog i drugog reda. U slučaju sistema prvog reda, prema definiciji kvantizacije zaokružavanjem (16.4) i (16.11) se dobija:

$$\left|Q\left(a\hat{y}[n-1]\right) - a\hat{y}[n-1]\right| \le 0.5 \cdot 2^{-B} \tag{16.123}$$

Kada se sistem nađe u graničnom ciklusu, očigledno važi:

$$\left|Q\left(a\hat{y}[n-1]\right)\right| = \left|\hat{y}[n-1]\right| \tag{16.124}$$

tj. efektivna vrednost koeficijenta a postaje jednaka jedinici. Opseg vrednosti koeficijenta a za koji je uslov (16.124) zadovoljen, određen je nejednačinom:

$$\left|\hat{y}[n-1]\right| - \left|a\hat{y}[n-1]\right| \le 0.5 \cdot 2^{-B} \tag{16.125}$$

čije je rešenje:

$$\left|\hat{y}[n-1]\right| \le \frac{0.5}{1-|a|} 2^{-B} = k \cdot 2^{-B} \tag{16.126}$$

Dakle, vrednosti izlaznog signala u mrtvoj zoni su multipli od 2^{-B}. Ako je $|a| < 0.5$, granični ciklus se ne može pojaviti. Ako je $|a| \ge 0.5$, po prestanku pobude izlazni signal počinje da opada i zaustavlja se na vrednosti $k \cdot 2^{-B}$, ili osciluje između vrednosti $k \cdot 2^{-B}$ i $-k \cdot 2^{-B}$, sa učestanošću jednakom polovini učestanosti odabiranja.

U slučaju sistema drugog reda, važi diferencna jednačina:

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + x[n] \tag{16.127}$$

kojoj odgovara funkcija prenosa sa kompleksnim polovima ako je $a_1^2 < -4a_2$. Sistem postaje nestabilan za $a_2 = 1$. Ako se kvantovanje proizvoda vrši pre sabiranja, diferencna jednačina (16.127) dobija oblik:

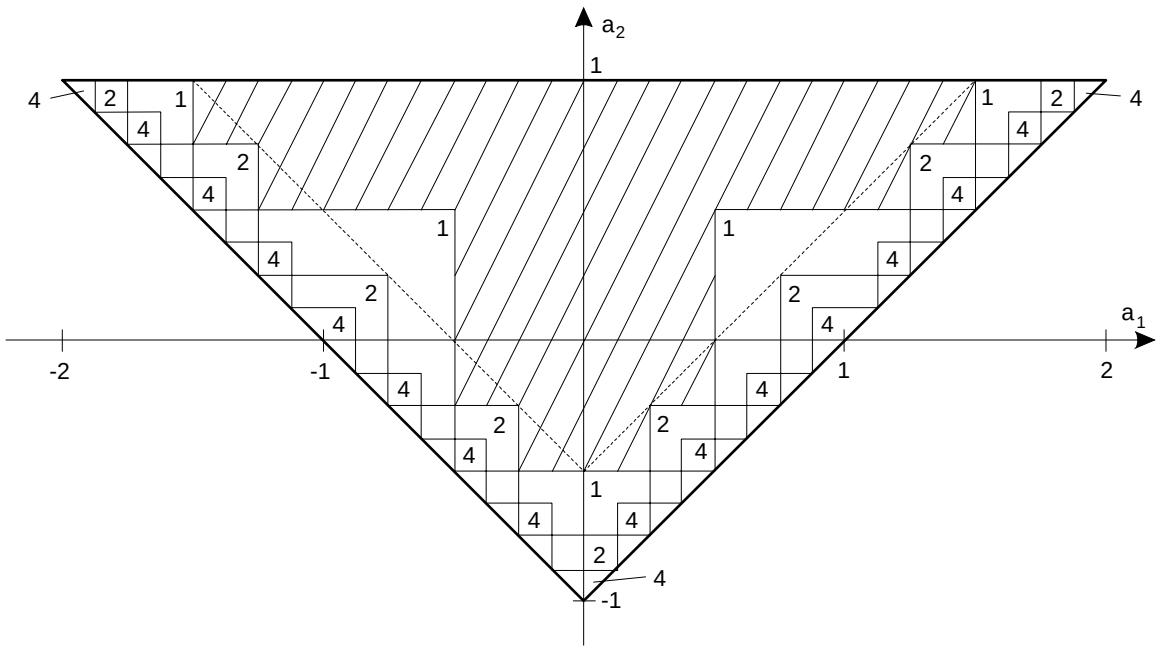
$$\hat{y}[n] = Q\left(a_1 \hat{y}[n-1]\right) + Q\left(a_2 \hat{y}[n-2]\right) + x[n] \tag{16.128}$$

Detaljnijom analizom moguće je odrediti uslove pod kojima može doći do oscilacija graničnog ciklusa kod sistema drugog reda. Pokazano je da kod sistema drugog reda postoje dva tipa graničnog ciklusa. U prvom slučaju, koji je potpuno analogan ponašanju sistema prvog reda, izlazni signal je konstantan ili osciluje između dve vrednosti sa učestanošću koja je jednaka polovini učestanosti odabiranja. Tada se iz jednačine (16.128), uz uslov $x[n] = 0$, za amplitudu oscilacija dobija jednačina:

$$y_{\max} = \pm Q\left(a_1 y_{\max}\right) - Q\left(a_2 y_{\max}\right) \tag{16.129}$$

Ovu nelinearnu jednačinu zadovoljava mnogo parova vrednosti a_1 i a_2, koje su prikazane na slici 16.13. Brojevi upisani u pojedine oblasti označavaju vrednosti amplitude oscilacija $y_{\max}$, koja se dobija za vrednosti para koeficijenata a_1 i a_2 iz te oblasti. Za vrednosti koeficijenata a_1 i a_2 iz šrafirane oblasti ne postoji ovaj tip graničnog ciklusa. Potreban, ali ne i dovoljan, uslov za egzistenciju graničnog ciklusa označen je isprekidanom linijom na slici 16.13 i glasi:

$$|a_1| \ge \frac{y_{\max} - 1}{y_{\max}} + a_2 \tag{16.130}$$



Slika 16.13 Oblasti koeficijenata funkcije drugog reda u kojima postoji prvi tip graničnog ciklusa.

U drugom slučaju, dobijaju se prave sinusoidalne oscilacije jer se konjugovano kompleksni polovi pomeraju na jedinični krug. Tada se, za kvantizaciju proizvoda zaokružavanjem, dobija:

$$\left|Q\left(a_2\hat{y}[n-2]\right)-a_2\hat{y}[n-2]\right| \le 0.5\cdot 2^{-B} \tag{16.131}$$

Polovi sistema naći će se na jediničnom krugu ako je:

$$Q\left(a_2\hat{y}[n-2]\right) = -\hat{y}[n-2] \tag{16.132}$$

tako da se zamenom u jednačinu (16.131) dobija:

$$\left|\hat{y}[n-2]\right| - \left|1+a_2\right| \le 0.5\cdot 2^{-B} \tag{16.133}$$

odnosno,

$$\left|\hat{y}[n-2]\right| \le \frac{0.5}{\left|1+a_2\right|}2^{-B} = k\cdot 2^{-B} \tag{16.134}$$

Dakle, ako je ulazni signal jednak nuli, a $\hat{y}[n-2]$ leži u opsegu određenom sa (16.134), polovi sistema drugog reda se pomeraju na jedinični krug. Učestanost oscilacija određena je koeficijentom a_1 i približno iznosi:

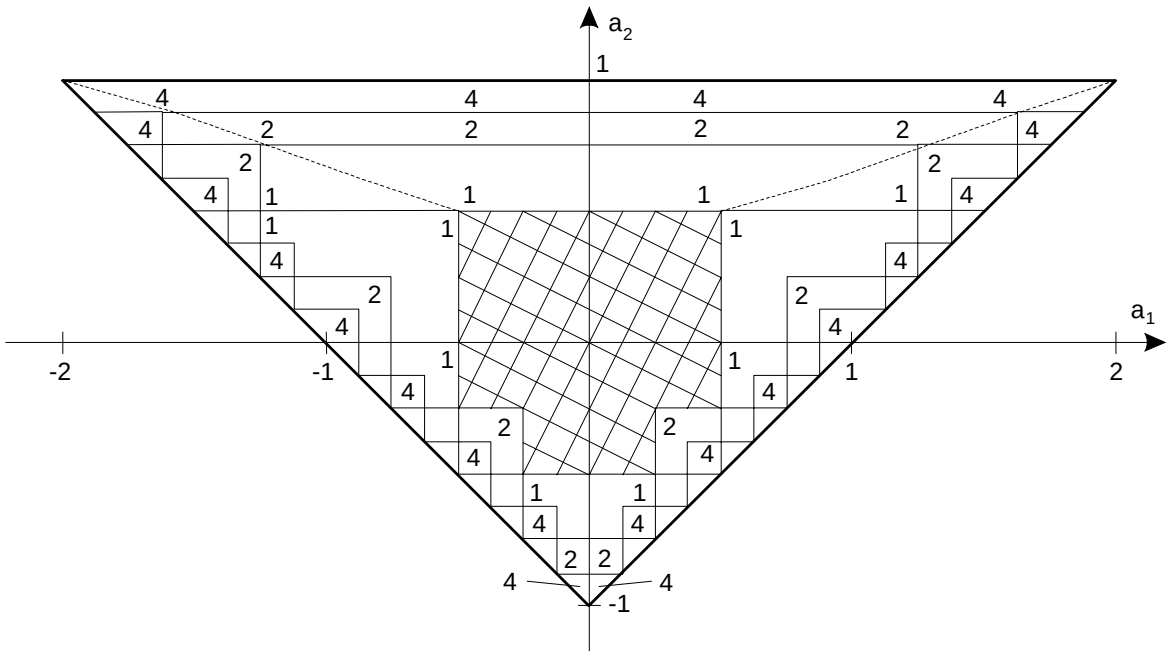
$$a_1 \approx -2\cos\Omega_o \tag{16.135}$$

gde je Ω_o racionalni umnožak učestanosti odabiranja. Na slici 16.14 u prostoru koeficijenata a_1 i a_2 prikazane su oblasti koje odgovaraju i ovom tipu oscilacija.

Prethodna analiza graničnog ciklusa izvedena je za slučaj kvantovanja sa zaokružavanjem. U literaturi je ispitivan i slučaj kvantovanja sa odsecanjem i pokazano je da se na taj način mogu eliminisati granični ciklusi u velikom broju slučajeva. Međutim, zbog toga što odsecanje proizvodi korelisani šum i veću snagu šuma, postupak odsecanja se retko primenjuje u praksi.

Granični ciklus se može u potpunosti izbeći podešavanjem vrednosti koeficijenata a_1 i a_2, ili izborom pogodne realizacione strukture. U svakom slučaju, amplituda graničnog ciklusa se može

smanjiti na prihvatljivi nivo povećanjem broja bita za reprezentaciju signala i koeficijenata. U literaturi se mogu naći brojne formule kojima se određuju granice amplitude oscilacija koje se dobro slažu sa rezultatima koji su dobijeni simulacijom. Dobijeni rezultati direktno su primenljivi i na paralelnu realizaciju funkcija prenosa višeg reda. Na žalost, kod kaskadne realizacije situacija je znatno komplikovanija, jer samo prva sekcija ima ulazni signal jednak nuli. Zbog toga se otpornost kaskadne realizacije na pojavu graničnog ciklusa najčešće ispituje simulacijom na računaru.



Slika 16.14 Oblasti koeficijenata funkcije drugog reda u kojima postoji prvi i drugi tip graničnog ciklusa.

U novije vreme, a naročito kod integrisanih digitalnih procesora signala, razvijena je još jedna tehnika za potiskivanje graničnog ciklusa usled kvantovanja proizvoda. To je tehnika *akumulatora dvostruke dužine*. Postupak se sastoji u tome da se parcijalne sume izračunavaju sa dvostrukom dužinom reči, a da se kvantovanje rezultata vrši tek kada se izračuna cela suma.

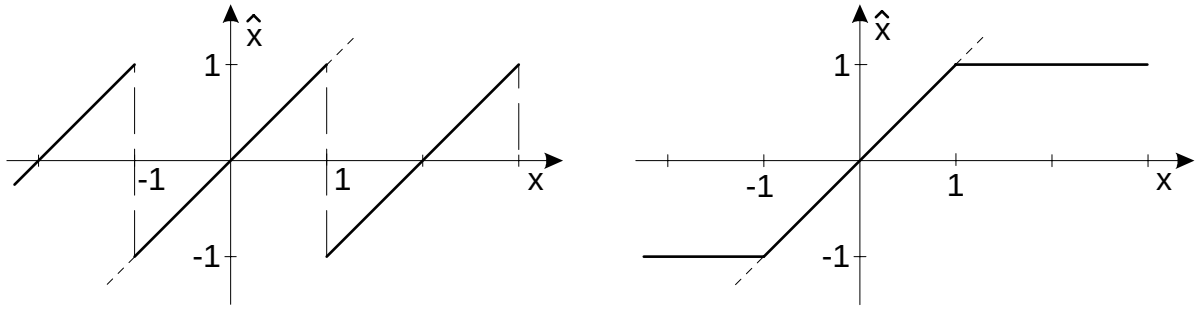
Granični ciklus kod struktura drugog reda je mnogo proučavan u literaturi zbog praktične važnosti ovog problema. To je posebno važno u primenama gde diskretni sistem neprekidno radi, tako da postoje vremenski intervali kada nema ulaznog signala. Postojanje graničnog ciklusa kod takvih sistema izazvalo bi nepoželjne izlazne signale koji mogu ozbiljno narušiti pravilno funkcionisanje sistema

Naravno, ima i primena gde je granični ciklus poželjna karakteristika sistema. Tipičan primer su digitalni sinusoidalni oscilatori kao i generatori koeficijenata za DFT.

16.5.2 GRANIČNI CIKLUSI ZBOG PREKORAČENJA OPSEGA PRI SABIRANJU

Druga vrsta nelinearnih efekata kod sistema za digitalnu obradu signala su granični ciklusi koji nastaju kao posledica prekoračenja opsega kod sabiranja. Kao što je već rečeno, u digitalnim sistemima za obradu signala se najčešće koristi komplement dvojke za predstavljanje bipolarnih signala i koeficijenata. Uobičajeno je da se svi signali i koeficijenti normalizuju na opseg $-1 \le x < 1$. I pored normalizacije može doći do prekoračenja opsega ako su oba sabirka istog znaka a po modulu su veći od 0.5. Ako se sa x označi rezultat običnog sabiranja a sa $\hat{x}$ rezultat sabiranja u sistemu komplementa dvojke, onda je karakteristika $\hat{x} = f(x)$ nelinearna i ima oblik prikazan na

slici 16.15a. Kao što se vidi, prilikom prekoračenja opsega dolazi do drastične greške u rezultatu, koja pokriva skoro ceo raspoloživi dinamički opseg. Zbog toga se takvo prekoračenje opsega mora sprečiti. Kao primer posmatrajmo ponovo diferencnu jednačinu drugog reda (16.127), odnosno (16.128), sa četvorobitnim koeficijentima prikazanim komplementom dvojke $a_1 = 0.75 = 0.110_2$ i $a_2 = -0.75 = 1.010_2$. Neka je ulazni signal $x[n] = 0$ za $n \geq 0$, $\hat{y}[-1] = 0.75 = 0.110_2$ i $\hat{y}[-2] = -0.75 = 1.010_2$. Ako se izračuna $\hat{y}[0]$, skraćivanjem proizvoda na dužinu od 4 bita i sabiranjem po pravilima za komplement dvojke, dobija se $\hat{y}[0] = 0.100100 + 0.100100 = 0.101 + 0.101 = 1.010 = -0.75$, umesto vrednosti $y[0] = 1.125$. Nastavljajući proces, dobija se $\hat{y}[1] = 1.011 + 1.011 = 0.110 = 0.75$. Dalje izračunavanje pokazuje da sistem osciluje između vrednosti 0.75 i -0.75 sa periodom koja je jednaka dvostrukoj periodi odabiranja. Oscilacije graničnog ciklusa ovde su posledica prekoračenja opsega pri sabiranju.



(a) **(b)**

Slika 16.15 Karakteristike sabiranja u komplementu dvojke: (a) standardna karakteristika, (b) karakteristika sa zasićenjem.

Detaljnije ispitivanje direktne realizacije sistema drugog reda pokazuje da je potreban i dovoljan uslov da se ne pojavi granični ciklus usled prekoračenja opsega pri sabiranju:

$$|a_1| + |a_2| < 1 \tag{16.136}$$

koji je vrlo strog i teško se može zadovoljiti u praksi. Stoga se obično pribegava drugom rešenju koje se sastoji u modifikaciji nelinearne karakteristike sabiranja sa slike 16.15a na način koji je prikazan na slici 16.15b. Takva karakteristika se naziva *sabiranje sa zasićenjem* i u potpunosti sprečava pojavu graničnog ciklusa u realizacijama funkcija prenosa drugog reda. Greška koju unosi ovakva karakteristika znatno je manja od greške koju unosi karakteristika sa slike 16.15a, pogotovu što se prekoračenje opsega retko javlja.

Na kraju treba reći da je pojava graničnog ciklusa usled kvantovanja proizvoda ili prekoračenja opsega sabirača moguća samo kod IIR sistema. Kod FIR sistema nije moguća pojava parazitnih oscilacija jer nema povratne sprege. Zbog toga se, u slučajevima kada se zahteva rad sistema bez graničnog ciklusa, često koriste i FIR sistemi.

16.6 UTICAJ KONAČNE DUŽINE REČI NA IZRAČUNAVANJE DFT

Efekti konačne dužine reči pojavljuju se i prilikom izračunavanja Diskretne Furijeove transformacije, naročito ako se to izračunavanje obavlja u aritmetici sa fiksnim položajem tačke. Međutim, zbog složenosti izračunavanja DFT, naročito kada se koriste neki efikasni algoritmi, analiza efekata konačne dužine reči na izračunavanje DFT je izuzetno složena. Zbog toga se kao i u

slučaju filtarskih struktura pribegava uprošćavanjima, koja se uglavnom sastoje od linearizacije nelinearnih efekata.

Efekti konačne dužine reči kod izračunavanja DFT manifestuju se na dva načina. Prvi uzrok grešaka, koji nastaje zbog nemogućnosti tačnog izračunavanja sinusa i kosinusa, sličan je efektu kvantovanja koeficijenata kod filtarskih struktura. Drugi uticaj konačne dužine reči nastaje zbog kvantovanja proizvoda i manifestuje se kao šum na izlazu.

16.6.1 UTICAJ KVANTOVANJA PROIZVODA NA DIREKTNO IZRAČUNAVANJE DFT

U četvrtom poglavlju definisan je izraz za izračunavanje DFT konačne kompleksne sekvence $x[n]$:

$$X[k]=\sum_{n=0}^{N-1}x[n]W_N^{kn},\quad k=0,1,\ldots,N-1 \tag{16.137}$$

gde je $W_N=e^{-j2\pi/N}$. Uobičajeno je da se kompleksno množenje izračunava pomoću 4 realna množenja. Ako se posle svakog množenja vrši kvantovanje, svako kompleksno množenje unosi 4 izvora šuma kvantovanja u blok dijagram DFT.

Izraz (16.137) sličan je izrazu za konvoluciju, tako da je ponašanje DFT u pogledu izlaznog šuma slično ponašanju FIR filtarske funkcije, koje je analizirano u odeljku 16.4.4. Osnovna razlika je što je broj izvora šuma četiri puta veći. Dakle, u slučaju kada se koristi aritmetika sa fiksnom tačkom i kvantovanje proizvoda vrši pre sabiranja, na osnovu izraza (16.119) za varijansu šuma kvantovanja na izlazu može se pisati:

$$\sigma_\varepsilon^2=4N\frac{q^2}{12}=N\frac{q^2}{3}=\frac{N}{3}2^{-2B} \tag{16.138}$$

Ako se pri izračunavanju kvantovanje vrši posle sabiranja svih proizvoda, tako što se koristi akumulatorski registar dvostruke dužine, onda je u jednačini (16.138), $N=1$.

Kao i kod realizacije FIR filtra korišćenjem aritmetike sa fiksnom tačkom, potrebno je voditi računa da ne dođe do prekoračenja dozvoljenog dinamičkog opsega. Iz jednačine (16.137) sledi:

$$\left|X[k]\right|\le\sum_{n=0}^{N-1}\left|x[n]\right|<N,\quad k=0,1,\ldots,N-1 \tag{16.139}$$

ako je ulazni signal normalizovan, tj. $\left|x[n]\right|<1$. Da se ne bi pojavilo prekoračenje opsega treba da bude $\left|X[k]\right|<1$, tako da je dovoljan uslov za sprečavanje prekoračenja:

$$\sum_{n=0}^{N-1}\left|x[n]\right|<1 \tag{16.140}$$

Dakle, za sigurno sprečavanje prekoračenja, dovoljno je podeliti ulazni signal sa N. Međutim, uslov (16.140) nije potreban i često je suviše restriktivan. Tipičan primer je sekvenca $x[n]=A\delta[n]$ čija je DFT, $X[k]=A,\ k=0,\ldots,N-1$. Ako se vrednost konstante A nalazi u opsegu $1/N<A<1$, nejednakost (16.140) nije zadovoljena, a ipak ne dolazi do prekoračenja opsega.

Posmatrajmo sada ulaznu sekvencu $x[n]$ koja predstavlja beli šum čije vrednosti, posle skaliranja, leže u opsegu $-1/N \le x(n) < 1/N$. Onda je varijansa ulazne sekvence:

$$\sigma_x^2 = \frac{(2/N)^2}{12} = \frac{1}{3N^2} \tag{16.141}$$

a varijansa izlaznih koeficijenata $X[k]$:

$$\sigma_X^2 = N\sigma_x^2 = \frac{1}{3N} \tag{16.142}$$

pa se za odnos signal-šum na izlazu dobija:

$$\frac{\sigma_X^2}{\sigma_\varepsilon^2} = \frac{1}{q^2N^2} = \frac{2^{2B}}{N^2} \tag{16.143}$$

Iz prethodnih jednačina se vidi da se skaliranjem smanjuje odnos signal-šum N puta, kao i da kombinacija skaliranja i grešaka kvantovanja smanjuje odnos signal-šum N^2 puta. Dakle, *skaliranjem ulaznog signala se rešava problem prekoračenja dinamičkog opsega po cenu ozbiljne redukcije odnosa signal-šum na izlazu*.

Kao primer, posmatrajmo sekvencu od 1024 odbirka. Ako je potrebno ostvariti odnos signal-šum od 30 dB, iz (16.143) se dobija da je potrebna tačnost množenja i sabiranja $B = 15$ bita. Stoga se ponekad odustaje od skaliranja ulazne sekvence već se samo zahteva da bude $|x[n]| < 1$. Tada se mora obezbediti dovoljno veliki dinamički opseg sabirača, jer je $|X[k]| < N$. Varijansa ulazne sekvence je $\sigma_x^2 = 1/3$, a varijansa izlazne sekvence $\sigma_X^2 = N\sigma_x^2 = N/3$. Tada je odnos signal-šum:

$$\frac{\sigma_X^2}{\sigma_\varepsilon^2} = \frac{1}{q^2} = 2^{2B} \tag{16.144}$$

odakle se vidi da je za $SNR = 30$ dB potrebno svega $B = 5$ bita. Naravno, potrebno je dodatnih 10 bita u akumulatorskom sabiraču da ne bi došlo do prekoračenja opsega. Prednost ovog rešenja je što množači rade samo sa 5 bita umesto sa 15 bita što može biti od značaja, naročito u hardverskim realizacijama.

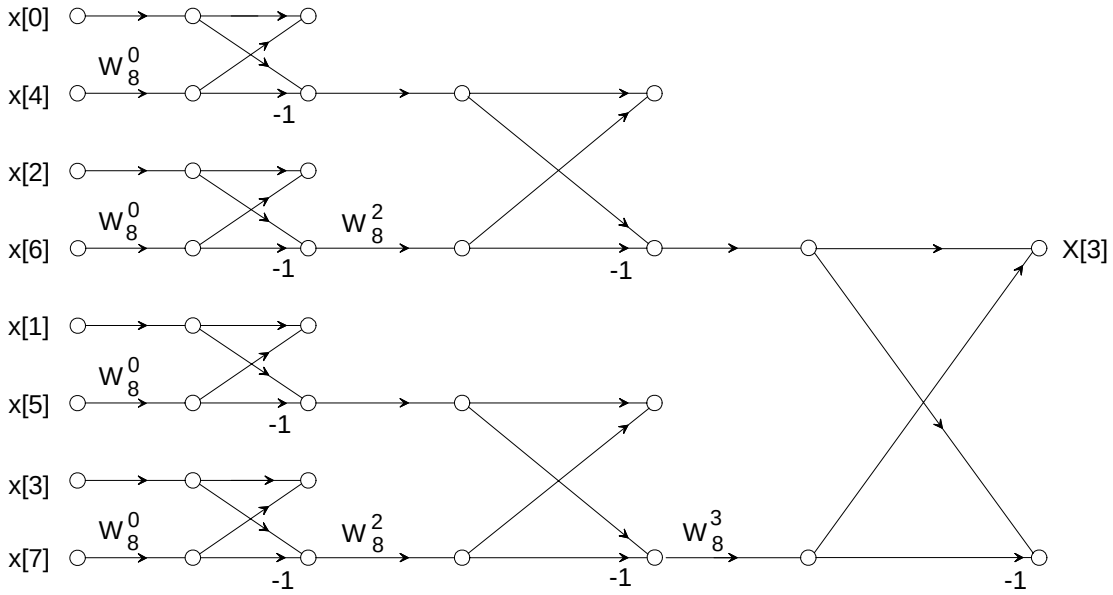
Dakle, vidi se da se problem prekoračenja opsega kod direktnog izračunavanja DFT može rešavati na razne načine. Prvo, ulazna sekvenca se može podeliti sa N, što pogoršava SNR na izlazu. Drugi način je da se za formiranje sume proizvoda koristi akumulator sa dodatnim brojem bita levo od tačke. Treći način je da se koristi blokovska aritmetika sa pokretnom tačkom, gde se vrši deljenje svih podataka sa 2 kad god se javi prekoračenje, a o kojoj će biti više reči kod analize FFT algoritama gde se češće koristi. Na kraju, može se koristiti aritmetika sa pokretnom tačkom, čime se problem prekoračenja opsega praktično eliminiše.

16.6.2 UTICAJ KVANTOVANJA PROIZVODA KOD FFT ALGORITAMA

Analizu uticaja kvantovanja proizvoda kod brzih algoritama za izračunavanje DFT teško je sprovesti u opštem slučaju, jer je struktura algoritama vrlo različita. Ipak, iz analize brzih algoritama sledi da se izračunavanje DFT za sekvencu velike dužine svodi na više izračunavanja DFT sekvenci manje dužine. Zbog toga je korisno proučiti efekte kvantovanja kod transformacija manje dužine, a dobijeni rezultati se mogu lako generalizovati.

Kao primer biće izvedena analiza efekata kvantovanja kod algoritma za brzo izračunavanje DFT preuređivanjem ulazne sekvence (DIT FFT). Dijagram toka DIT FFT algoritma prikazan je na slici 5.6, odakle se vidi da se u svakom stepenu izračunava vektor od N elemenata na osnovu vektora od N elemenata iz prethodnog stepena. Elementi vektora se izračunavaju u parovima, pomoću leptir operacije. Za izračunavanje svakog od izlaznih odbiraka potrebno je izračunati $N/2$ leptira u prvom stepenu, $N/4$ u drugom, $N/8$ u trećem, itd., dok se u poslednjem stepenu izračunava samo jedan leptir. Kao primer, na slici 16.16 je prikazan postupak izračunavanja odbirka $X[3]$ kada je dužina ulazne sekvence $N = 8$. Dakle, broj leptira koje je potrebno izračunati za jedan izlazni odbirak jednak je:

$$1 + 2 + 4 + \cdots + N/8 + N/4 + N/2 = N - 1 \tag{16.145}$$



Slika 16.16 Uticaj kvantovanja proizvoda na izračunavanje DFT odbirka X(3).

Pošto u svakom leptiru postoji samo jedno kompleksno množenje, odnosno 4 realna množenja, ukupan broj množenja potreban za izračunavanje jednog izlaznog odbirka je $4(N-1)$. Položaj množača u dijagramu toka je različit, tako da greške propagiraju na razne načine do izlaza. Međutim, rotacioni faktori kojima se množe signali u dijagramu toka ne utiču na statističke karakteristike grešaka kvantovanja jer imaju jediničnu amplitudu. Pod pretpostavkom da su izvori šuma kvantovanja nekorelisani, varijansa ukupne greške kvantovanja na izlazu je:

$$\sigma_\varepsilon^2 = 4(N-1)\frac{q^2}{12} \approx N\frac{q^2}{3} = \frac{N}{3}2^{-2B} \tag{16.146}$$

Interesantno je da se dobija *isti rezultat kao kod direktnog izračunavanja DFT*. Ovaj rezultat je prvi pogled neočekivan, zbog toga što direktno izračunavanje DFT zahteva znatno više množenja nego FFT algoritmi. Objašnjenje ove pojave je jednostavno. Za izračunavanje jednog izlaznog odbirka, i kod direktnog i kod brzog izračunavanja, potreban je isti broj množenja, a ušteda u broju operacija kod FFT algoritama potiče od toga što se pojedina množenja višestruko koriste.

Kao što je već uočeno kod izračunavanja DFT pomoću direktnog algoritma, problem prekoračenja opsega može biti vrlo ozbiljan kada je broj odbiraka, odnosno broj stepena izračunavanja, veliki. Ako se izvrši skaliranje ulaznog signala deljenjem svakog odbirka sa N,

prekoračenje opsega se sigurno sprečava jer je $|X[k]|<1$. U tom slučaju važe izrazi (16.141), (16.142) i (16.143) koji su izvedeni za direktno izračunavanje DFT kada ulazna sekvenca predstavlja beli šum. Kod takvog načina skaliranja odnos signal-šum opada sa N^2, odnosno *1 bit po stepenu*. To znači da je potrebno povećati dužinu reči za 1 bit, ako se broj odbiraka poveća dva puta, da bi se dobio isti odnos signal-šum na izlazu. Ovaj rezultat važi i za ulazne signale koji ne predstavljaju beli šum.

Korišćenje aritmetičke jedinice dvostruke dužine, koje je vrlo efikasno smanjivalo uticaj šuma usled kvantovanja proizvoda kod realizacija filtarskih funkcija i kod direktnog izračunavanja DFT, malo pomaže kod FFT algoritama. Naime, pošto se izračunavanje obavlja po stepenima, posle izračunavanja svakog leptira potrebno je dobijene rezultate smestiti u memoriju, a tada se rezultat obavezno mora skratiti na normalnu dužinu reči. Dakle, uvođenje aritmetičke jedinice dvostruke dužine smanjuje greške kvantovanja samo tokom izračunavanja leptir operacije.

Kod brzih algoritama mogući su i drugi načini skaliranja koji daju bolje rezultate u pogledu odnosa signal-šum. Posmatrajmo dva sukcesivna stepena izračunavanja, $m-1$ i m, kod algoritma sa preuređivanjem u vremenu. Za leptire u m-tom stepenu važe jednačine:

$$\begin{aligned} X_m[p] &= X_{m-1}[p] + W_N^r X_{m-1}[q] \\ X_m[q] &= X_{m-1}[p] - W_N^r X_{m-1}[q] \end{aligned} \tag{16.147}$$

Iz jednačina (16.147) se lako dobijaju sledeće nejednakosti:

$$\max\left(|X_{m-1}[p]|,|X_{m-1}[q]|\right) \le \max\left(|X_m[p]|,|X_m[q]|\right) \tag{16.148}$$

$$\max\left(|X_m[p]|,|X_m[q]|\right) \le 2\max\left(|X_{m-1}[p]|,|X_{m-1}[q]|\right) \tag{16.149}$$

Koristeći ove relacije, ukupan faktor skaliranja $1/N$ se može raspodeliti na pojedine stepene FFT algoritma. Na primer, ako je $|x[n]|<1$, onda se u prvom stepenu može izvršiti skaliranje sa 0.5 tako da bude $x[n]<0.5$. Ako se posle toga u svakom stepenu vrši skaliranje sa 0.5, ukupan faktor skaliranja biće $1/2^p = 1/N$. Na ovaj način se takođe potpuno eliminiše pojava prekoračenja opsega u FFT algoritmu. Ovakav postupak skaliranja ne menja amplitudu izlaznih odbiraka u odnosu na postupak skaliranja deljenjem ulaznih odbiraka sa N, ali znatno smanjuje varijansu šuma kvantovanja na izlazu. Naime, svako množenje sa 0.5 smanjuje varijansu greške kvantovanja 4 puta. Zbog toga se greške kvantovanja iz prvog stepena kojih ima $4(N/2)$ smanjuju 4^{p-1} puta, greške iz drugog stepena kojih ima $4(N/4)$ smanjuju se 4^{p-2} puta, itd. Ukupna varijansa greške kvantovanja na izlazu FFT algoritma je:

$$\begin{aligned} \sigma_\varepsilon^2 &= \frac{q^2}{12}\left\{4\frac{N}{2}\frac{1}{4^{p-1}} + 4\frac{N}{4}\frac{1}{4^{p-2}} + 4\frac{N}{8}\frac{1}{4^{p-3}} + \cdots + 4\right\} \\ &= \frac{q^2}{3}\left\{\frac{1}{2^{p-1}} + \frac{1}{2^{p-2}} + \frac{1}{2^{p-3}} + \cdots + \frac{1}{2} + 1\right\} = \frac{2q^2}{3}\left(1 - \frac{1}{2^p}\right) \approx \frac{2}{3}2^{-2B} \end{aligned} \tag{16.150}$$

Vidi se da se ovakvim načinom skaliranja gubi linearna zavisnost varijanse šuma na izlazu od broja odbiraka N, kao u izrazima (16.138) i (16.146). Zbog toga je odnos signal-šum:

$$\frac{\sigma_X^2}{\sigma_\varepsilon^2} = \frac{1}{2N} 2^{2B} = 2^{2B-p-1} \tag{16.151}$$

odnosno, odnos *signal-šum je inverzno proporcionalan sa N* umesto sa N^2. Drugim rečima, za zadati odnos signal-šum, svaki novi stepen izračunavanja kod FFT algoritma zahteva produženje digitalne reči za 0.5 bita.

Odredimo sada potreban broj bita za izračunavanje DFT sekvence od 1024 odbiraka, ako se traži odnos signal-šum od 30 dB. Na osnovu (16.151) se dobija $B = 11$, što je za 4 bita manje nego kada se skaliranje koncentriše u prvom stepenu FFT algoritma.

Treći pristup postupku skaliranja radi sprečavanja prekoračenja opsega je korišćenje *blokovske aritmetike sa pokretnom tačkom*. Originalni odbirci se tako normalizuju da imaju maksimalne mantise koje obezbeđuju da bude $|x[n]| < 1$. Svi odbirci imaju isti eksponent. Postupak izračunavanja leptirova u prvom i narednim stepenima izvodi se aritmetikom sa fiksnom tačkom uz dodatno testiranje prekoračenja opsega posle svakog sabiranja. Kada se otkrije prekoračenje, celi vektor se podeli sa 2 i postupak se nastavlja. Broj deljenja sa 2 se pamti i određuje ukupni faktor skaliranja na izlazu. Odnos signal-šum kod blokovske aritmetike sa pokretnom tačkom je u proseku bolji za oko 3 dB nego u prethodno opisanom postupku raspodeljenog skaliranja, ali jako zavisi od ulaznog signala.

16.6.3 UTICAJ KVANTOVANJA KOD FFT ALGORITAMA I ARITMETIKE SA POKRETNOM TAČKOM

Iz analize izvedene u prethodnom odeljku može se izvesti važan zaključak da je prekoračenje opsega najvažniji faktor, koji utiče na odnos signal-šum kod implementacije FFT algoritama koja koristi aritmetiku sa fiksnom tačkom. Kako se prekoračenje opsega praktično ne može pojaviti ako se koristi aritmetika sa pokretnom tačkom, logično je očekivati da su karakteristike FFT algoritama implementiranih pomoću aritmetike sa pokretnom tačkom bolje. U brojnim istraživanjima koja su se bavila ovom problematikom pokazano je da se korišćenjem aritmetike sa pokretnom tačkom izbegava skaliranje. Pokazano je da je *za aritmetiku sa pokretnom tačkom odnos signal-šum inverzno proporcionalan sa* $p = \log_2 N$ umesto sa N kao što je to bio slučaj kada se koristi aritmetika sa fiksnom tačkom i raspodeljeno skaliranje. Dakle, za četiri puta veće p (sekvenca dužine N^4) i isti odnos signal-šum, potreban je samo 1 dodatni bit. Ovaj rezultat je eksperimentalno proveren simulacijom na računaru.

16.6.4 UTICAJ KVANTOVANJA KOEFICIJENATA KOD FFT ALGORITAMA

U prethodnoj analizi uticaja konačne dužine reči kod algoritama za izračunavanje DFT bilo je pretpostavljeno da su vrednosti rotacionih faktora apsolutno tačne, što naravno nije tačno ako se koristi konačan broj bita za predstavljanje sinusa i kosinusa. Mada je priroda grešaka usled kvantovanja koeficijenata sama po sebi nestatistička, vrlo korisni rezultati se mogu dobiti statističkom analizom. Analiza se sastoji u tome da se svaki koeficijent predstavi svojom tačnom vrednošću kojoj se dodaje aditivni beli šum koji predstavlja kvantovanje. Na ovaj način se dobija rezultat koji pokazuje da je odnos signal-šum na izlazu približno jednak $(6/p)2^{2B}$. Iz ovog rezultata se može izvesti važan zaključak da odnos signal-šum usled kvantovanja koeficijenata opada vrlo sporo sa porastom dužine sekvence N, tačnije opada sa faktorom $p = \log_2 N$. Ovaj zaključak je i

eksperimentalno potvrđen pri čemu je uočeno da su praktični rezultati uvek bolji od teorijskih predviđanja.

Korišćenje aritmetike sa pokretnom tačkom donosi dalje poboljšanje odnosa signal-šum usled kvantovanja koeficijenata za približno 4 puta.

16.7 RAČUNARSKA SIMULACIJA KVANTOVANJA

U prethodnim analizama izvedeni su izrazi koji pokazuju uticaj kvantovanja koeficijenata i kvantovanja proizvoda na izlazni signal. Međutim, u praksi se često sreću sistemi na koje se izvedeni rezultati ne mogu direktno primeniti jer se po strukturi ili načinu izvođenja aritmetičkih operacija razlikuju od sistema za koji su izvedeni prethodni rezultati. Osim toga, veliki broj izvedenih rezultata je statističke prirode, što znači da važe samo sa određenim stepenom sigurnosti. Zbog toga se često primenjuje ispitivanje sistema simulacijom na računaru pre nego što se sistem implementira na raspoloživom hardveru.

Računarska simulacija aritmetike sa fiksnom tačkom izvodi se pisanjem odgovarajućeg programa koji simulira izvođenje operacija sabiranja i množenja u sistemu sa fiksnom tačkom. Program se obično piše na nekom višem programskom jeziku. U suštini, potrebno je napisati samo jedan potprogram za kvantovanje sa zaokružavanjem ili odsecanjem, ali je, radi jednostavnije realizacije celog programa za simulaciju, pogodno napisati i posebne potprograme za sabiranje i množenje.

Potprogram za simulaciju kvantovanja zasnovan je na sledećim principima. Neka je x broj sa pokretnom tačkom i brojem bita koji je podržan od strane korišćenog programskog jezika, koji predstavlja tačnu vrednost koeficijenta ili signala u sistemu. Vrednost broja x treba konvertovati u drugi broj $\hat{x}$, koji je u sistemu komplementa dvojke sa fiksnom tačkom po vrednosti najbliži broju x. Inače, interna predstava broja $\hat{x}$ u računaru ista je kao predstava broja x. Neka broj $\hat{x}$ simulira broj sa fiksnom tačkom koji ima NI bita levo od tačke koji predstavljaju celobrojni deo broja uključujući znak i NF bita desno od tačke koji predstavljaju razlomački deo broja. U praksi se najviše koriste vrednosti $NI = 1$ i $NF = 15,\ 23$ ili 31. Na ovaj način se mogu simulirati brojevi sa fiksnom tačkom iz opsega $-2^{NI-1} \le \hat{x} < 2^{NI-1} - 2^{-NF}$ sa rezolucijom 2^{-NF}.

U FORTRAN-u je najjednostavnije simulirati kvantovanje korišćenjem sistemske funkcije `IFIX` koja vrši konverziju realnih brojeva u celobrojne postupkom odsecanja. Postupak zaokruživanja može se simulirati dodavanjem 0.5 pozitivnom broju i oduzimanjem 0.5 od negativnog broja, pre nego što se izvrši konverzija funkcijom `IFIX`. Da bi se izvršilo korektno kvantovanje brojeva koji imaju razlomački deo, potrebno je prvo broj pomnožiti nekom konstantom $QK = 2^{NF}$, izvršiti konverziju u celi broj funkcijom `IFIX`, a posle toga rezultat podeliti sa QK. Potprogram za simulaciju kvantovanja `QUANT` prikazan je na slici 16.17.

Da bi se olakšala simulacija procesa kvantovanja i izbeglo pisanje dugih programa, pogodno je napisati posebne potprograme za sabiranje i množenje sa kvantovanjem rezultata koji koriste opisani potprogram `QUANT`. Potprogrami za sabiranje (`QADD`) i množenje (`QMULT`) su prikazani na slikama 16.18 i 16.19 i zbog jednostavnosti ne zahtevaju posebna objašnjenja.

U programu za simulaciju rada realnog digitalnog sistema za obradu signala prvo se svi koeficijenti kvantuju potprogramom `QUANT`. Zatim se u petlji (ili petljama) u kojoj se izračunava suma proizvoda svako množenje zameni pozivom potprograma `QMULT`, a svako sabiranje pozivom

potprograma QADD. Ukoliko se simulira aritmetička jedinica koja koristi akumulator dvostruke dužine, onda se osnovni program menja samo dodavanjem poziva za potprogram QUANT posle izračunavanja sume proizvoda, odnosno kada se simulira smeštanje međurezultata u memorijsku lokaciju ili registar standardne tačnosti.

```
C
C   POTPROGRAM ZA SIMULACIJU KVANTOVANJA
C   X  -  BROJ KOJI SE KVANTUJE
C   NI -  BROJ BITA ZA CELOBROJNI DEO I ZNAK
C   NF -  BROJ BITA ZA RAZLOMACKI DEO
C   RNDOFF  =  TRUE  - KVANTOVANJE ZAOKRUZAVANJEM
C           =  FALSE - KVANTOVANJE ODSECANJEM
C
      FUNCTION QUANT (X, NI, NF, RNDOFF)
      LOGICAL RNDOFF
C
C   PROVERA DA LI BROJ X LEZI U DOZVOLJENOM OPSEGU
C
      XMIN = -2.**(NI-1)
      XMAX = -XMIN - 2.**(-NF)
      IF (X .LE. XMIN) THEN
         QUANT = XMIN
         RETURN
      ENDIF
      IF (X .GE. XMAX) THEN
         QUANT = XMAX
         RETURN
      ENDIF
C
C   KVANTOVANJE
C
      IF (RNDOFF) THEN
         RND = 0.5D0
      ELSE
         RND = 0.D0
      ENDIF
      QK = 2.**NF
      IF (X .GE. 0.D0) THEN
         QUANT = IFIX (QK*X + RND)/QK
      ELSE
         QUANT = IFIX (QK*X - RND)/QK
      ENDIF
      RETURN
      END
```

Slika 16.17 Potprogram za simulaciju kvantovanja QUANT.

```
C
C   POTPROGRAM ZA SIMULACIJU SABIRANJA SA KVANTOVANJEM
C   TERM1 -  PRVI SABIRAK
C   TERM2 -  DRUGI SABIRAK
C   NI    -  BROJ BITA ZA CELOBROJNI DEO I ZNAK
C   NF    -  BROJ BITA ZA RAZLOMACKI DEO
C   RNDOFF  =  TRUE  - KVANTOVANJE ZAOKRUZAVANJEM
C           =  FALSE - KVANTOVANJE ODSECANJEM
C
      FUNCTION QADD (TERM1, TERM2, NI, NF, RNDOFF)
      LOGICAL RNDOFF
      ADD = TERM1 + TERM2
      QADD = QUANT (ADD, NI, NF, RNDOFF)
      RETURN
      END
```

Slika 16.18 Potprogram za simulaciju sabiranja sa kvantovanjem.

```
C
C   POTPROGRAM ZA SIMULACIJU MNOZENJA SA KVANTOVANJEM
C   TERM1 -  PRVI CINILAC
C   TERM2 -  DRUGI CINILAC
C   NI    -  BROJ BITA ZA CELOBROJNI DEO I ZNAK
C   NF    -  BROJ BITA ZA RAZLOMACKI DEO
C   RNDOFF  =  TRUE  -  KVANTOVANJE ZAOKRUZAVANJEM
C           =  FALSE - KVANTOVANJE ODSECANJEM
C
      FUNCTION QMULT (TERM1, TERM2, NI, NF, RNDOFF)
      LOGICAL RNDOFF
      RMULT = TERM1 * TERM2
      QMULT = QUANT (RMULT, NI, NF, RNDOFF)
      RETURN
      END
```

Slika 16.19 Potprogram za simulaciju množenja sa kvantovanjem.