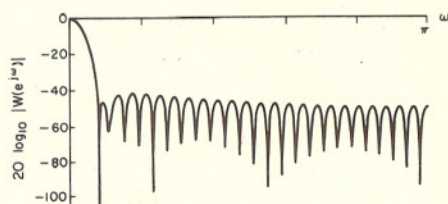


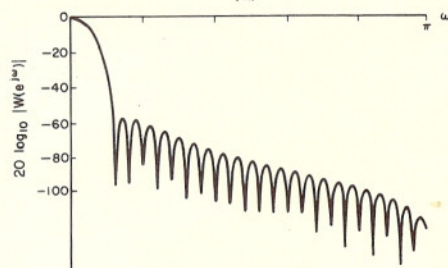
A.V. Oppenheim
R.W. Schafer

ELABORAZIONE NUMERICA DEI SEGNALI

Edizione italiana a cura di
C. Braccini e G. Gambardella



(d)



(e)

Ingegneria elettrica

FrancoAngeli

In copertina : la foto rappresenta la struttura discreta a simmetria circolare, proposta da G. Sandini e V. Tagliasco dell'Istituto di Elettrotecnica della Università di Genova, per campionare ed elaborare immagini con risoluzione spaziale linearmente decrescente con l'eccentricità (in analogia con il sistema visivo umano).

10^a edizione: 1994

Titolo originale: *Digital Signal Processing*

Copyright © 1975 by Prentice-Hall, Inc., Englewood Cliffs, N.J., Usa

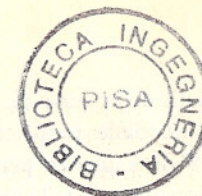
Unica traduzione italiana autorizzata di Carlo Braccini e Giuseppe Gambardella

Copyright © by FrancoAngeli s.r.l., Milano, Italy.

È vietata la riproduzione, anche parziale o ad uso interno o didattico, con qualsiasi mezzo effettuata, non autorizzata. Stampa Tipomonta, viale Monza 126, Milano.

I lettori che desiderano essere regolarmente informati sulle novità pubblicate dalla nostra Casa Editrice possono scrivere, mandando il loro indirizzo, alla "FrancoAngeli, Viale Monza 106, 20127 Milano", ordinando poi i volumi direttamente alla loro Libreria.

INDICE

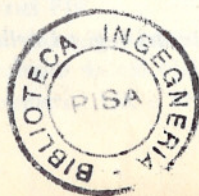


Prefazione all'edizione italiana, di Carlo Braccini e Giuseppe Gambardella	pag. 9
Prefazione	» 11
Introduzione	» 17
1. Segnali e sistemi a tempo discreto	» 23
1.0 Introduzione	» 23
1.1 Segnali a tempo discreto - Sequenze	» 25
1.2 Sistemi lineari invarianti alla traslazione	» 28
1.3 Stabilità e causalità	» 33
1.4 Equazioni lineari alle differenze a coefficienti costanti	» 34
1.5 Rappresentazione nel dominio della frequenza di sistemi e segnali a tempo discreto	» 37
1.6 Alcune proprietà di simmetria della trasformata di Fourier	» 43
1.7 Campionamento di segnali a tempo continuo	» 45
1.8 Sequenze e sistemi bidimensionali	» 50
Sommario	» 54
Problemi	» 56
SI 2. La trasformata z	» 66
2.0 Introduzione	» 66
2.1 La trasformata z	» 66
2.2 La trasformata z inversa	» 74
2.3 Teoremi e proprietà della trasformata z	» 79
2.4 Funzione di trasferimento	» 90
WD 2.5 La trasformata z bidimensionale	» 96
Sommario	» 101
Problemi	» 102
SI 3. La trasformata di Fourier discreta	» 111
3.0 Introduzione	» 111
3.1 Rappresentazione di sequenze periodiche - La serie di Fourier discreta	» 112

3.2 Proprietà della serie di Fourier discreta	pag. 116
3.3 Riassunto delle proprietà della rappresentazione con la DFS di sequenze periodiche	» 120
3.4 Campionamento della trasformata z	» 120
3.5 Rappresentazione di Fourier per sequenze di durata finita - La trasformata di Fourier discreta	» 123
3.6 Proprietà della trasformata di Fourier discreta	» 126
3.7 Riassunto delle proprietà della trasformata di Fourier discreta	» 136
3.8 Convoluzione lineare basata sulla trasformata di Fourier discreta	» 136
NO 3.9 Trasformata di Fourier discreta bidimensionale	» 143
Sommario	» 148
Problemi	» 148
4. <i>Uso dei grafi di flusso e matrici per la rappresentazione dei filtri numerici</i>	» 163
4.0 Introduzione	» 163
SI 4.1 Rappresentazione delle reti numeriche mediante grafi di flusso di segnale	» 164
NO 4.2 Rappresentazione matriciale delle reti numeriche	» 171
SI 4.3 Le strutture di rete fondamentali per sistemi IIR	» 177
SI 4.4 Forme trasposte	» 182
SI 4.5 Le strutture di rete fondamentali per sistemi FIR	» 185
NO 4.6 Effetti della quantizzazione dei parametri	» 195
NO 4.7 Il teorema di Tellegen per i filtri numerici e sue applicazioni	» 205
Sommario	» 213
Problemi	» 215
5. <i>Tecniche di progetto di filtri numerici</i>	» 228
5.0 Introduzione	» 228
5.1 Progetto di filtri numerici IIR da filtri analogici	» 230
5.2 Esempi di progetto: trasformazione analogico-numerica	» 244
5.3 Progetto di filtri numerici IIR assistito da calcolatore	» 264
5.4 Proprietà dei filtri numerici FIR	» 270
5.5 Progetto di filtri FIR con l'uso di finestre	» 272
5.6 Progetto di filtri FIR assistito da calcolatore	» 284
5.7 Un confronto tra filtri numerici IIR e FIR	» 302

Sommario	pag. 303
Problemi	» 305
6. <i>Calcolo della trasformata di Fourier discreta</i>	» 317
6.0 Introduzione	» 317
6.1 L'algoritmo di Goertzel	» 320
6.2 Algoritmi di FFT basati sulla decimazione nel tempo	» 322
6.3 Algoritmi di FFT basati sulla decimazione in frequenza	» 334
6.4 Algoritmi di FFT per N numero composto	» 340
6.5 Considerazioni generali su problemi di calcolo per gli algoritmi di FFT	» 347
6.6 Algoritmo della trasformata z chirp	» 353
Sommario	» 358
Problemi	» 361
7. <i>Trasformate di Hilbert discrete</i>	» 369
7.0 Introduzione	» 369
7.1 Sufficienza della sola parte reale o immaginaria per sequenze causali	» 371
7.2 Condizione di fase minima	» 377
7.3 Trasformate di Hilbert per la DFT	» 385
7.4 Trasformate di Hilbert per sequenze complesse	» 389
Sommario	» 397
Problemi	» 399
8. <i>Segnali casuali discreti</i>	» 407
8.0 Introduzione	» 407
8.1 Un processo casuale a tempo discreto	» 408
8.2 Medie	» 413
8.3 Rappresentazioni in frequenza dei segnali a energia infinita	» 419
8.4 Risposta dei sistemi lineari a segnali casuali	» 422
Sommario	» 426
Problemi	» 427
9. <i>Effetti della lunghezza finita dei registri nella elaborazione numerica dei segnali</i>	» 434
9.0 Introduzione	» 434

9.1 Effetto della rappresentazione dei numeri sulla quantizzazione	pag. 436
9.2 La quantizzazione nel campionamento di segnali analogici	» 443
9.3 Effetti della lunghezza finita dei registri nella realizzazione di filtri numerici IIR	» 448
9.4 Effetti della lunghezza finita dei registri nelle realizzazioni di filtri numerici FIR	» 469
9.5 Effetti della lunghezza finita dei registri nei calcoli della trasformata di Fourier discreta	» 475
Sommario	» 493
Problemi	» 495
10. <i>Elaborazione omomorfa dei segnali</i>	» 510
10.0 Introduzione	» 510
10.1 Sovrapposizione generalizzata	» 511
10.2 Sistemi omomorfi moltiplicativi	» 513
10.3 Elaborazione omomorfa di immagini	» 516
10.4 Sistemi omomorfi per la convoluzione	» 520
10.5 Proprietà del cepstrum complesso	» 530
10.6 Algoritmi per la realizzazione del sistema caratteristico D_*	» 536
10.7 Applicazioni della deconvoluzione omomorfa	» 541
Sommario	» 557
Problemi	» 558
11. <i>Stima dello spettro di potenza</i>	» 561
11.0 Introduzione	» 561
11.1 Principi fondamentali di teoria della stima	» 562
11.2 Stime dell'autocovarianza	» 567
11.3 Il periodogramma come stima dello spettro di potenza	» 570
11.4 Stimatori dello spettro « smussati »	» 576
11.5 Stima della covarianza incrociata e dello spettro incrociato	» 583
11.6 Uso della FFT nella stima dello spettro	» 584
11.7 Esempio di stima dello spettro	» 591
Sommario	» 600
Problemi	» 600
Indice analitico	» 605



PREFAZIONE ALL'EDIZIONE ITALIANA

L'edizione italiana del libro di Oppenheim e Schafer potrà essere utile a quanti, soprattutto a livello universitario, hanno bisogno di una preparazione di base nel campo della elaborazione numerica dei segnali per poter usare a questo scopo, in modo consapevole e creativo, i calcolatori e i vari altri dispositivi numerici che la tecnologia elettronica rende oggi disponibili. Ci sembra che proprio la crescente diffusione e il miglioramento nelle prestazioni (cioè costo, velocità, dimensioni) di tale tecnologia sottolineino il valore formativo di un libro come questo, che non è soggetto a « invecchiamento », in quanto è dedicato ai fondamenti dei segnali e sistemi discreti e non a particolari tecnologie realizzative. Non a caso il libro è stato già adottato con successo in più di cento università degli Stati Uniti. C'è da dire piuttosto che il libro non pretende di coprire l'intera area della teoria dei segnali e sistemi discreti e che, da quando è stato scritto, si sono pur verificati degli interessanti sviluppi in alcuni settori dell'elaborazione numerica dei segnali. A questi, che possiamo considerare limiti del libro, e che meglio potrebbero chiamarsi delimitazioni, accenneremo ora brevemente, anche perché non ci sembra utile sottolinearne ulteriormente i pregi. A parte approfondimenti teorici in argomenti particolari, un fatto importante di questi ultimi anni è stato l'avvicinamento delle comunicazioni e dei controlli in settori come l'identificazione e la stima. Da questo punto di vista, mancano nel libro un'introduzione alla rappresentazione dei sistemi discreti mediante variabili di stato e tutto l'argomento del filtraggio ottimo e della stima lineare di parametri, inclusa la stima dello spettro mediante modelli autoregressivi. Altri sviluppi, successivi alla stesura di questo libro, riguardano le tecniche di elaborazione in parallelo e il campo cui esse particolarmente si prestano, cioè il trattamento delle immagini. Inoltre, si può citare ancora l'uso di trasformate diverse da quella di Fourier nella rappresentazione dei segnali, e i corrispondenti algoritmi di trasformazione veloce.

Qualche parola, infine, su alcune scelte riguardanti la traduzione. Innanzitutto si è deciso di rendere l'inglese *digital* non con un neologismo (« digitale ») ma con l'aggettivo « numerico » che è di significato sostanzialmente uguale e di uso diffuso nella letteratura tecnica. Tale scelta è stata mantenuta anche per dizioni non proprio felici (come in « rete numerica »), dove comunque il contesto è più che sufficiente a chiarire il significato.

INTRODUZIONE

L'elaborazione numerica* dei segnali, una disciplina che affonda le sue radici nelle matematiche del XVII e XVIII secolo, è diventata oggi un importante strumento di lavoro in una quantità di settori diversi della scienza e della tecnologia. Le tecniche e le applicazioni di questa disciplina, mentre da un lato risalgono a Newton e a Gauss, sono al tempo stesso intimamente connesse ai moderni calcolatori numerici e ai circuiti integrati.

L'elaborazione numerica dei segnali si occupa della rappresentazione dei segnali mediante sequenze di numeri o simboli e della elaborazione di tali sequenze. Lo scopo di questa elaborazione può essere quello di stimare dei parametri caratteristici di un segnale o quello di trasformare un segnale in una forma che per qualche motivo risulta più vantaggiosa. Le tecniche dell'analisi numerica classica, come quelle che servono per l'interpolazione, l'integrazione e la derivazione, sono certamente algoritmi dell'elaborazione numerica dei segnali. D'altra parte, la disponibilità di calcolatori numerici molto veloci ha incoraggiato lo sviluppo di algoritmi di elaborazione sempre più complessi e sofisticati, e inoltre i recenti progressi nella tecnologia dei circuiti integrati fanno prevedere realizzazioni economiche di sistemi molto complessi di elaborazione numerica dei segnali.

L'elaborazione dei segnali, in generale, ha una ricca storia, e la sua importanza è evidente in campi molto diversi come l'ingegneria biomedica, l'acustica, le tecniche sonar e radar, la sismologia, la comunicazione del parlato e quella dei dati, la fisica nucleare, e molti altri campi. In diverse applicazioni, come, per esempio, nelle analisi elettroencefalografiche (EEG) ed elettrocardiografiche (ECG), o in sistemi per la trasmissione o il riconoscimento del parlato, l'obiettivo può essere quello di estrarre alcuni parametri caratteristici del segnale. In alternativa, lo scopo può essere quello di separare il segnale utile da altri segnali che interferiscono (« rumore »), oppure quello di modificare il segnale per presentarlo in una

* Per la traduzione dell'aggettivo *digital* del testo americano è stato preferito, qui e nel seguito, il termine « numerico » al neologismo « digitale », in quanto è adeguato nel significato e più comunemente usato nella letteratura tecnica in italiano (n.d.t.).

forma che sia più facilmente interpretabile da un esperto. Come ulteriore esempio, si consideri il caso di un segnale trasmesso attraverso un canale di telecomunicazioni. Questo segnale risulta generalmente perturbato in una quantità di modi che includono la distorsione del canale, l'attenuazione e l'inserimento del rumore di fondo. Uno degli obiettivi del ricevitore è quello di compensare tutti questi disturbi. In ogni caso, si richiede una elaborazione del segnale di ricezione.

I problemi di elaborazione dei segnali non si limitano, ovviamente, al caso di segnali monodimensionali. Molte applicazioni di elaborazione delle immagini richiedono l'uso di tecniche di elaborazione di segnali bidimensionali. Valgano come esempi i problemi dell'analisi e del miglioramento (« enhancement ») di radiografie, di fotografie aeree per la rivelazione di incendi di foreste, danni a raccolti, o condizioni meteorologiche, e di trasmissioni televisive da sonde lunari o nel lontano spazio. Tecniche di elaborazione di segnali multidimensionali sono anche utilizzate per l'analisi di dati sismici in relazione a terremoti, ricerche di giacimenti petroliferi e controllo di esperimenti nucleari.

Fino a pochi anni fa l'elaborazione dei segnali è stata attuata usando tipicamente strumentazione analogica. Alcune eccezioni si ebbero già negli anni '50, specialmente là dove si richiedeva un'elaborazione particolarmente sofisticata. Questo è accaduto, ad esempio, per l'analisi di certi dati geofisici che venivano registrati su nastro magnetico per poi elaborarli su grandi calcolatori numerici. Questo tipo di problemi costituisce uno dei primi esempi di trattamento di segnali usando i calcolatori numerici. Questa elaborazione non poteva generalmente essere svolta in tempo reale; per esempio occorrevano minuti, o anche ore, di tempo al calcolatore per elaborare talora solo pochi secondi di dati. Pur con questa limitazione, la flessibilità del calcolatore numerico rendeva questa alternativa assai attraente.

Durante questo stesso periodo si sviluppò anche un altro modo di usare i calcolatori numerici nella elaborazione dei segnali. Grazie alla flessibilità dei calcolatori numerici si dimostrò infatti utile in molti casi simulare un sistema di elaborazione di segnali su calcolatore numerico, prima di passare alla sua realizzazione con circuiti analogici. In questo modo un nuovo algoritmo, o sistema, di elaborazione poteva essere studiato sperimentalmente e in modo flessibile, prima di impegnare risorse economiche e tecniche nella sua costruzione. Esempi tipici sono state le simulazioni dei « vocoder » sviluppate presso il « Lincoln Laboratory » e i « Bell Laboratories ». Nella realizzazione di un « vocoder » a canali analogico, le caratteristiche dei filtri spesso influiscono in modo non prevedibile sulla qualità del segnale voce risultante. Ebbene, queste caratteristiche dei filtri venivano messe a punto attraverso simulazioni su calcolatore e la qualità del sistema poteva essere valutata prima della costruzione del dispositivo analogico.

In tutti gli esempi citati sopra di elaborazione di segnali fatta mediante calcolatori numerici, il calcolatore offriva indubbiamente enormi vantaggi di flessibilità. Tuttavia non sempre l'elaborazione poteva essere svolta in tempo reale. Di conseguenza l'atteggiamento prevalentemente sviluppatosi in quegli anni è stato quello di usare il calcolatore numerico per *approssimare*, o *simulare*, un sistema analogico di elaborazione di segnali. Questo atteggiamento ha fatto sì che i primi lavori sul filtraggio numerico riguardassero soprattutto i diversi modi di programmare un filtro in modo che la sequenza convertitore analogico/numerico-filtro numerico-convertitore numerico/analogico approssimasse un buon filtro analogico. L'idea, infatti, che i sistemi numerici potessero essere usati praticamente per la effettiva e diretta elaborazione dei segnali (per la voce, il radar o altre applicazioni) sembrava allora molto azzardata. Velocità, costo e dimensioni erano, ovviamente, tre importanti fattori a favore dell'uso di componenti analogici.

Man mano che i segnali venivano elaborati su calcolatori numerici, si sviluppava tuttavia una naturale tendenza a sperimentare algoritmi di elaborazione di segnali sempre più sofisticati. Alcuni di questi algoritmi nacquero come prodotti diretti della flessibilità del calcolatore numerico e non sembravano avere alcuna possibilità di realizzazione pratica nella strumentazione analogica. Così molti di questi algoritmi erano considerati interessanti, ma, in un certo senso, poco utili. Un esempio di una classe di algoritmi di questo tipo fu l'insieme delle tecniche indicate col nome di « cepstrum » e filtraggio omomorfo. Era stato dimostrato in modo inequivocabile su calcolatore che queste tecniche potevano essere vantaggiosamente applicate per sistemi di compressione di banda nel parlato, sistemi per la deconvoluzione e per la rimozione di echi. La messa in opera di queste tecniche richiede il calcolo esplicito della trasformata inversa di Fourier del logaritmo della trasformata di Fourier dell'ingresso. La precisione e la risoluzione necessarie nel calcolo della trasformata di Fourier erano tali da escludere l'uso degli analizzatori di spettro analogici. Ostacoli di questo tipo fornivano quindi un'ulteriore spinta alla realizzazione in modo del tutto numerico di sistemi di elaborazione dei segnali. Cominciò pertanto un lavoro attivo di ricerca su « vocoder » e analizzatori di spettro numerici ed altri sistemi completamente numerici, nella speranza che alla fine tali sistemi sarebbero diventati pratici e convenienti.

Lo sviluppo di un nuovo modo di considerare l'elaborazione numerica dei segnali fu anche accelerato dalla scoperta, nel 1965, di un algoritmo efficiente per il calcolo delle trasformate di Fourier. Questa classe di algoritmi è conosciuta con il nome di trasformata di Fourier veloce o FFT (« Fast Fourier Transform »). Le sue implicazioni si dimostrarono importanti da diversi punti di vista. Numerosi algoritmi di elaborazione dei segnali che erano stati sviluppati su calcolatori numerici richiedevano

tempi di elaborazione che superavano di parecchi ordini di grandezza il tempo reale. Spesso ciò dipendeva dal fatto che l'analisi spettrale era una parte importante dell'elaborazione e non si conosceva nessun mezzo efficiente per realizzarla. La trasformata di Fourier veloce veniva a ridurre il tempo di calcolo della trasformata di Fourier di diversi ordini di grandezza. Ciò permise la realizzazione di algoritmi sempre più sofisticati per il trattamento dei segnali con tempi di elaborazione che consentivano l'interattività con il sistema. Inoltre, avendo compreso che l'algoritmo della trasformata di Fourier veloce poteva in effetti realizzarsi mediante circuiti ad esso specificamente dedicati, molti algoritmi di elaborazione dei segnali che fino ad allora erano sembrati « non pratici » cominciarono a potersi realizzare grazie proprio a circuiti (logici) specializzati.

Un'altra importante implicazione dell'algoritmo della trasformata di Fourier veloce dipendeva dal suo essere fin dall'origine un concetto a tempo discreto. Esso, infatti, mirava al calcolo della trasformata di Fourier di un segnale a tempo discreto ovvero di una sequenza, ed implicava proprietà e formalismi matematici esatti nel dominio del tempo discreto. Non si trattava dunque di una semplice approssimazione della trasformata di Fourier nel dominio del tempo continuo. Ciò si dimostrò importante perché ne seguì uno stimolo a riformulare molti concetti e algoritmi di elaborazione dei segnali nel formalismo adatto al tempo discreto, di modo che queste tecniche formarono un insieme esatto di relazioni nel nuovo dominio. Ci si allontanava così dall'idea che l'elaborazione dei segnali su un calcolatore numerico fosse semplicemente un'approssimazione delle tecniche di elaborazione dei segnali analogici. Con questo cambiamento di punto di vista si affermava un forte interesse nel nuovo, o rinato, campo dell'elaborazione numerica dei segnali.

Le tecniche e le applicazioni dell'elaborazione numerica dei segnali vanno espandendosi ad un ritmo straordinario. Con l'avvento dell'integrazione su larga scala e la conseguente riduzione nei costi e nelle dimensioni dei componenti (insieme alla aumentata velocità), la classe delle applicazioni delle tecniche di elaborazione numerica si allarga continuamente. È oggi possibile realizzare filtri numerici specializzati funzionanti a velocità di campionamento nella fascia dei megahertz. Sono anche disponibili in commercio elaboratori specializzati che realizzano la trasformata di Fourier veloce, per elevati flussi di dati. Semplici filtri numerici sono stati realizzati con un unico circuito integrato. Per quanto riguarda i sistemi di compressione di banda del parlato, l'orientamento attuale è pressoché unanimemente diretto verso realizzazioni tutte numeriche grazie alla loro attuale maggiore praticità. Gli elaboratori numerici si trovano anche ad essere parte integrante di numerosi sistemi moderni di radar e sonar. C'è da aggiungere che, oltre ai circuiti specializzati per l'elaborazione numerica dei segnali, sono oggi disponibili speciali calcolatori programmabili

la cui architettura è specificamente adattata a problemi di elaborazione numerica di segnali. Tali calcolatori stanno trovando applicazioni sia nell'elaborazione dei segnali in tempo reale, sia per simulazioni in tempo reale orientate allo sviluppo di circuiti logici specializzati.

L'importanza dell'elaborazione numerica dei segnali sembra crescere continuamente senza nessun segno evidente di saturazione. In effetti lo sviluppo futuro di questo campo si annuncia ancora più straordinario di quello che abbiamo appena descritto. L'impatto delle tecniche di elaborazione numerica dei segnali promuoverà senza dubbio eccezionali progressi in alcuni campi di applicazione. Un esempio notevole è il campo della telefonia, dove le tecniche numeriche fanno intravedere un vero e proprio salto di economia e flessibilità nella realizzazione dei sistemi di commutazione e trasmissione.

Proprio per la direzione dell'evoluzione di questo campo, noi siamo convinti che le tecniche di elaborazione numerica dei segnali debbano essere studiate e considerate per se stesse piuttosto che come un'approssimazione all'elaborazione analogica. In questo libro noi partiamo dall'ipotesi che il lettore abbia familiarità con la rappresentazione dei segnali e la teoria dei sistemi lineari a tempo continuo. Cominciamo pertanto il primo capitolo definendo l'insieme dei segnali e sistemi di cui ci occupiamo e di lì sviluppiamo le nostre tecniche. La maggior parte della nostra enfasi è sulla classe dei sistemi lineari invarianti alla traslazione, che corrispondono ai sistemi lineari tempo-invarianti del caso continuo. Nel capitolo 10, tuttavia, prenderemo in considerazione una generalizzazione di questa classe di sistemi.

Nella presentazione degli argomenti abbiamo fatto uno sforzo consapevole per evitare di introdurre forzatamente nel quadro dell'elaborazione numerica dei segnali concetti che sono propri dell'elaborazione analogica. Risulterà tuttavia evidente fin dai primi capitoli che molte delle idee e dei risultati relativi all'elaborazione numerica dei segnali hanno il loro corrispondente nell'elaborazione analogica. La convoluzione, per esempio, è uno strumento importante nella rappresentazione dei sistemi lineari invarianti alla traslazione. Analogamente un ruolo essenziale è giocato dal dominio delle frequenze e dall'analisi di Fourier. Ma, oltre alle forti somiglianze, fra le impostazioni analogica e numerica esistono anche forti e importanti differenze. Per questo, benché le conoscenze preesistenti di elaborazione analogica possano a volte tornare utili, vogliamo mettere in guardia il lettore dal lasciarsi fuorviare da tali conoscenze e intuizioni nella comprensione dell'elaborazione numerica dei segnali.

Sebbene l'elaborazione numerica dei segnali sia un campo di conoscenze dinamico e in rapida espansione, si può dire che le sue basi sono ben formulate. Molte di queste basi hanno origine nell'analisi numerica classica sviluppata nel 1600. Importanti rifiniture delle tecniche fonda-

mentali dell'elaborazione numerica dei segnali si devono allo sviluppo e al trattamento dei sistemi di controllo a dati campionati degli anni 1940 e 1950, ed altre allo sviluppo dell'elaborazione numerica dei segnali nella sua forma recente. Gli argomenti presentati in questo libro sono stati scelti in modo da mettere il lettore in grado di sviluppare una solida base nei fondamenti di questo campo e cominciare ad apprezzarne l'ampio spettro di applicazioni e le tendenze future.

1. SEGNALI E SISTEMI A TEMPO DISCRETO

1.0 INTRODUZIONE

Un *segnale* può essere definito come una funzione o una grandezza che contiene informazione, in generale riguardo allo stato o al comportamento di un sistema fisico. Anche se i segnali possono essere rappresentati in molti modi, l'informazione è sempre contenuta nelle variazioni di una o più grandezze in qualche dominio. Per esempio, il segnale può essere costituito dall'insieme delle variazioni di una grandezza nel tempo o nello spazio.

Matematicamente i segnali sono rappresentati come funzioni di una o più variabili indipendenti. Per esempio, un segnale vocale può essere rappresentato matematicamente come una funzione del tempo ed una fotografia può essere rappresentata come una luminosità funzione di due variabili spaziali. È convenzione diffusa, e sarà seguita in questo libro, considerare il tempo come la variabile indipendente della rappresentazione matematica di un segnale, anche se in realtà tale variabile può non essere il tempo.

La variabile indipendente della rappresentazione matematica di un segnale può essere continua o discreta. Segnali *a tempo continuo* sono segnali definiti su un insieme temporale continuo e perciò sono rappresentati da funzioni di variabile continua. Segnali *a tempo discreto* sono quelli definiti su un insieme discreto di tempi e quindi la variabile indipendente assume solo valori discreti; in altri termini, i segnali a tempo discreto sono rappresentati come sequenze di numeri. Come vedremo in seguito, segnali come il parlato o le immagini possono avere una rappresentazione in termini di variabile continua oppure di variabile discreta e, se valgono certe condizioni, queste rappresentazioni sono del tutto equivalenti.

Oltre al fatto che le variabili indipendenti possono essere continue o discrete, anche il valore del segnale può essere continuo o discreto. Segnali *numerici* sono quelli per cui sia il tempo che l'ampiezza sono discreti. I segnali a tempo continuo e ad ampiezza continua sono a volte chiamati *segnali analogici*.

In quasi tutti i settori della scienza e della tecnologia occorre elaborare dei segnali per facilitare l'estrazione di informazione. Perciò lo sviluppo di tecniche e sistemi di elaborazione dei segnali è di grande importanza. Di solito queste tecniche assumono la forma di una trasformazione del segnale in un altro segnale che, per qualche motivo, risulta più vantaggioso dell'originale. Per esempio, si possono cercare trasformazioni che separano due o più segnali che sono stati combinati in qualche modo; si può desiderare di rendere più evidente qualche componente o qualche parametro di un segnale; oppure possiamo voler stimare uno o più parametri di un segnale. I sistemi di elaborazione dei segnali possono essere classificati allo stesso modo dei segnali. Così, *sistemi a tempo continuo* sono sistemi per cui sia l'ingresso che l'uscita sono segnali a tempo continuo e *sistemi a tempo discreto* sono quelli per cui l'ingresso e l'uscita sono segnali a tempo discreto. Analogamente, *sistemi analogici* sono sistemi per cui l'ingresso e l'uscita sono segnali analogici e *sistemi numerici* sono quelli per cui l'ingresso e l'uscita sono segnali numerici. L'*elaborazione numerica dei segnali* riguarda quindi le trasformazioni di segnali che sono discreti sia in ampiezza che in tempo. Questo capitolo, e in pratica la maggior parte di questo libro, tratta segnali e sistemi a tempo discreto invece che numerici. Gli effetti dell'ampiezza discreta sono esaminati in dettaglio nel cap. 9.

I segnali a tempo discreto possono avere origine dal campionamento di un segnale a tempo continuo oppure possono essere generati direttamente da qualche processo a tempo discreto. Qualunque sia l'origine dei segnali a tempo discreto, i sistemi di elaborazione numerica dei segnali presentano molti aspetti vantaggiosi. Essi possono essere realizzati in maniera molto flessibile usando calcolatori numerici di impiego generale (« general-purpose »), oppure possono essere realizzati con dispositivi e circuiti (« hardware ») numerici. Se necessario, essi possono essere usati per simulare sistemi analogici o, cosa più importante, per realizzare trasformazioni di segnali impossibili a farsi con circuiti analogici. Perciò le rappresentazioni numeriche dei segnali sono spesso necessarie quando è richiesta un'elaborazione sofisticata.

In questo capitolo considereremo i concetti fondamentali relativi ai segnali e sistemi di elaborazione a tempo discreto, dapprima per segnali monodimensionali e poi per segnali bidimensionali. L'enfasi sarà posta soprattutto sulla classe dei sistemi lineari a tempo discreto invarianti alla traslazione (« shift-invariant »). Vale per questo capitolo e per i successivi che molte delle proprietà e dei risultati che deriveremo saranno simili alle proprietà e ai risultati validi per i sistemi lineari tempo-invarianti a tempo continuo descritti in molti ottimi testi [1-3]. Effettivamente è possibile affrontare la trattazione dei sistemi a tempo discreto considerando le sequenze come segnali analogici costituiti da treni di impulsi. Questo

approccio, se accuratamente sviluppato, può portare a risultati corretti, tanto è vero che costituisce la base di gran parte della trattazione classica dei sistemi a dati campionati (v. per es. [4-6]). D'altro canto, in molte applicazioni attuali dell'elaborazione numerica dei segnali, non tutte le sequenze derivano dal campionamento di un segnale a tempo continuo. Per di più, molti sistemi a tempo discreto non sono semplicemente approssimazioni di corrispondenti sistemi analogici. Perciò, invece di cercare di adattare forzatamente al caso discreto risultati presi dalla teoria dei sistemi analogici, deriveremo risultati simili operando direttamente nel dominio dei sistemi a tempo discreto e usando la notazione ad essi adatta. I segnali a tempo discreto saranno posti in relazione con quelli analogici solo quando necessario.

1.1 SEGNALI A TEMPO DISCRETO - SEQUENZE

Nella teoria dei sistemi a tempo discreto occorre elaborare segnali che sono rappresentati da sequenze. Una sequenza di numeri x , in cui l' n .mo numero della sequenza è indicato con $x(n)$, è scritta formalmente come

$$x = \{x(n)\}, \quad -\infty < n < \infty \quad (1.1)$$

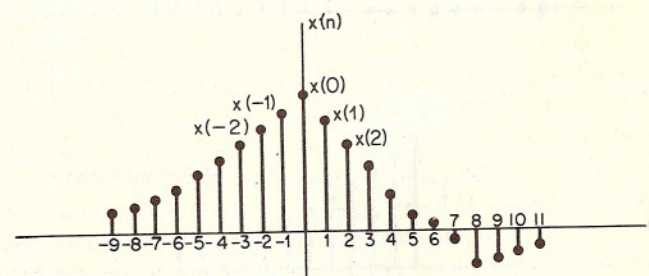


Fig. 1.1 Rappresentazione grafica di un segnale a tempo discreto.

Anche se le sequenze non sempre provengono dal campionamento di forme d'onda analogiche, per comodità chiameremo $x(n)$ il « campione n .mo » della sequenza. Inoltre, anche se in senso stretto $x(n)$ indica l' n .mo numero nella sequenza, la notazione della (1.1) è spesso sovrabbondante, per cui risulta comodo e non ambiguo usare l'espressione « la sequenza $x(n)$ ». I segnali a tempo discreto (cioè le sequenze) sono spesso rappresentati graficamente come in fig. 1.1. Anche se l'ascissa è disegnata come

una linea continua, è importante riconoscere che $x(n)$ è definita solo per valori interi di n . Non è corretto pensare che $x(n)$ sia zero per n non intero; $x(n)$ è semplicemente non definita per valori non interi di n .

In fig. 1.2 sono mostrati alcuni esempi di sequenze. La *sequenza campione unitario*, $\delta(n)$, è definita come la sequenza con valori

$$\delta(n) = \begin{cases} 0, & n \neq 0 \\ 1, & n = 0 \end{cases}$$

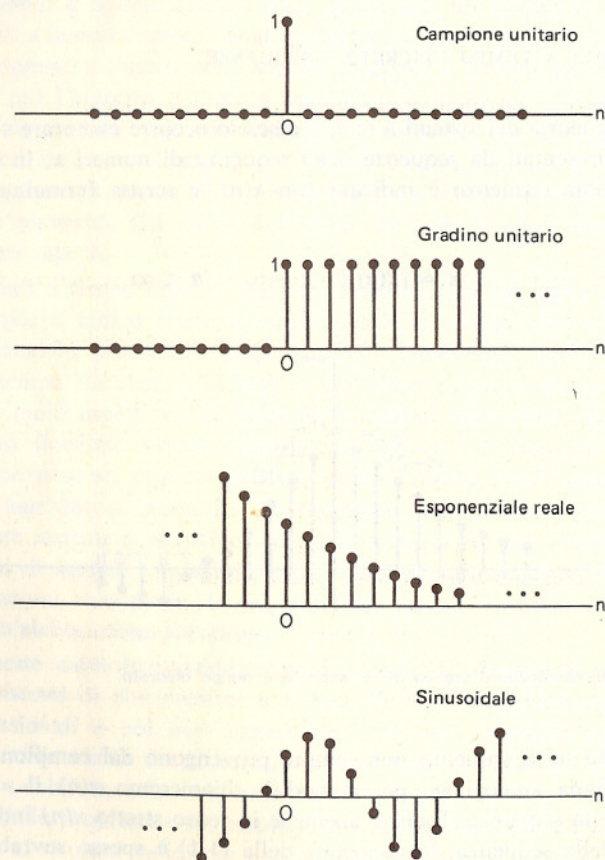


Fig. 1.2 Alcuni esempi di sequenze. Le sequenze rappresentate giocano un ruolo importante nell'analisi e nella rappresentazione di segnali e sistemi a tempo discreto.

Come vedremo tra poco, la sequenza campione unitario gioca per i segnali e sistemi a tempo discreto lo stesso ruolo che gioca la funzione impulsiva per i segnali e sistemi a tempo continuo. Per comodità, la sequenza campione unitario è spesso indicata come *impulso a tempo discreto* o semplicemente come *impulso*. È importante notare che l'impulso a tempo discreto non presenta le complicazioni matematiche dell'impulso a tempo continuo: la sua definizione è semplice e precisa. La sequenza gradino unitario, $u(n)$, ha valori

$$u(n) = \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

Il gradino unitario è legato al campione unitario da

$$u(n) = \sum_{k=-\infty}^n \delta(k) \quad (1.2)$$

Analogamente, il campione unitario può essere legato al gradino unitario da

$$\delta(n) = u(n) - u(n-1) \quad (1.3)$$

Una sequenza esponenziale reale è qualsiasi sequenza i cui valori siano della forma a^n , dove a è un numero reale. Una sequenza sinusoidale ha valori della forma $A \cos(\omega_0 n + \phi)$.

Una sequenza esponenziale complessa è della forma $e^{(\sigma + j\omega_0)n}$.

Una sequenza $x(n)$ si definisce *periodica* con periodo N se $x(n) = x(n + N)$ per ogni n . L'esponenziale complesso con $\sigma = 0$ e le sequenze sinusoidali hanno un periodo di $2\pi/\omega_0$ solo quando questo numero reale è un intero. Se $2\pi/\omega_0$ non è un intero ma un numero razionale, la sequenza sinusoidale sarà periodica ma con un periodo più lungo di $2\pi/\omega_0$. Se $2\pi/\omega_0$ non è un numero razionale, le sequenze sinusoidale ed esponenziale complessa non sono più periodiche. Il parametro ω_0 sarà considerato come frequenza della sinusoidale o dell'esponenziale complesso, siano queste periodiche o meno. La frequenza ω_0 può essere scelta in un insieme continuo di valori. Non si perde tuttavia in generalità vincolando ω_0 ad essere continua nel campo $0 \leq \omega_0 \leq 2\pi$ (o in quello equivalente $-\pi \leq \omega_0 \leq \pi$) in quanto le sequenze sinusoidale o esponenziale complessa ottenute facendo variare ω_0 nel campo $2\pi k \leq \omega_0 \leq 2\pi(k+1)$ sono del tutto identiche, per qualsiasi k , a quelle ottenute facendo variare ω_0 nell'intervallo $0 \leq \omega_0 \leq 2\pi$.

È a volte utile introdurre l'*energia* di una sequenza. L'energia \mathcal{E} di una sequenza $x(n)$ è definita come

$$\mathcal{E} = \sum_{n=-\infty}^{\infty} |x(n)|^2$$

Nell'analisi dei sistemi di elaborazione di segnali a tempo discreto le sequenze sono soggette ad alcune operazioni di base. Il prodotto e la somma di due sequenze x ed y si definiscono come, rispettivamente, la somma e il prodotto campione per campione:

$$x \cdot y = \{x(n)y(n)\} \quad (\text{prodotto})$$

$$x + y = \{x(n) + y(n)\} \quad (\text{somma})$$

La moltiplicazione di una sequenza x per un numero α è definita come

$$\alpha \cdot x = \{\alpha x(n)\}$$

Si dice che la sequenza y è una versione ritardata o traslata della sequenza x se y ha valori

$$y(n) = x(n - n_0)$$

dove n_0 è un intero.

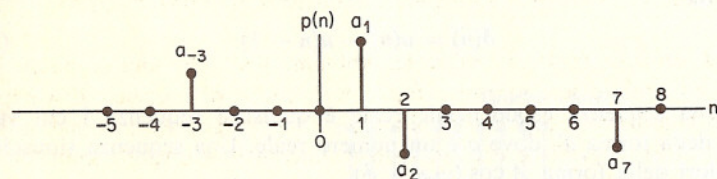


Fig. 1.3 Esempio di una sequenza espressa come la somma di campioni unitari scalati e ritardati.

Una sequenza arbitraria può essere espressa come la somma di campioni unitari scalati e ritardati. Per esempio, la sequenza $p(n)$ di fig. 1.3 può essere espressa come

$$p(n) = a_{-3} \delta(n+3) + a_1 \delta(n-1) + a_2 \delta(n-2) + a_7 \delta(n-7)$$

Più in generale, una qualsiasi sequenza può essere espressa come

$$x(n) = \sum_{k=-\infty}^{\infty} x(k) \delta(n-k) \quad (1.4)$$

1.2 SISTEMI LINEARI INVARIANTI ALLA TRASLAZIONE

Un sistema è definito matematicamente come una trasformazione univoca o un operatore che mappa una sequenza di ingresso $x(n)$ in una sequenza di uscita $y(n)$. Questo viene indicato con

$$y(n) = T[x(n)]$$

ed è spesso rappresentato come in fig. 1.4.

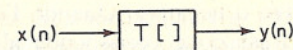


Fig. 1.4 Rappresentazione di una trasformazione che mappa una sequenza di ingresso $x(n)$ in una sequenza di uscita $y(n)$.

Imponendo dei vincoli sulla trasformazione $T[]$ si possono definire delle classi di sistemi a tempo discreto. Sarà studiata in dettaglio la classe dei sistemi lineari invarianti alla traslazione, poiché tali sistemi possono essere caratterizzati in maniera relativamente semplice dal punto di vista matematico e in quanto possono essere progettati per eseguire funzioni utili di elaborazione di segnali. Nel cap. 10 discuteremo una classe più generale di sistemi, di cui i sistemi lineari sono un caso particolare.

La classe dei sistemi lineari è definita mediante il principio di sovrapposizione. Se $y_1(n)$ e $y_2(n)$ sono le risposte rispettivamente agli ingressi $x_1(n)$ e $x_2(n)$, un sistema è lineare se e solo se

$$T[ax_1(n) + bx_2(n)] = aT[x_1(n)] + bT[x_2(n)] = ay_1(n) + by_2(n) \quad (1.5)$$

dove a e b sono costanti arbitrarie. Abbiamo visto che una sequenza arbitraria $x(n)$ può essere rappresentata come somma di sequenze di campioni unitari ritardati e scalati come nella (1.4). Questa rappresentazione, insieme con la (1.5), suggerisce che un sistema lineare può essere completamente caratterizzato dalla sua risposta al campione unitario*. In particolare, si supponga che $h_k(n)$ sia la risposta del sistema a $\delta(n-k)$, cioè ad un campione unitario che si manifesta quando $n = k$. Allora dalla (1.4)

$$y(n) = T\left[\sum_{k=-\infty}^{\infty} x(k) \delta(n-k)\right]$$

e usando la (1.5) possiamo scrivere

$$y(n) = \sum_{k=-\infty}^{\infty} x(k) T[\delta(n-k)] = \sum_{k=-\infty}^{\infty} x(k) h_k(n) \quad (1.6)$$

La risposta del sistema può dunque essere espressa, secondo la relazione (1.6), mediante la risposta del sistema a $\delta(n-k)$. Se si impone soltanto la linearità, $h_k(n)$ dipenderà sia da n che da k , e in questo caso la (1.6) è di utilità limitata. Si ottiene un risultato più utile se imponiamo anche il vincolo dell'invarianza alla traslazione.

La classe dei sistemi invarianti alla traslazione è caratterizzata dalla proprietà che, se $y(n)$ è la risposta a $x(n)$, $y(n-k)$ è la risposta a $x(n-k)$,

* Nel seguito useremo indifferentemente le dizioni « risposta al campione unitario » e « risposta all'impulso » (n.d.t.).

dove k è un intero positivo o negativo. Quando l'indice n è associato col tempo, l'invarianza alla traslazione corrisponde all'*invarianza nel tempo*. Dalla proprietà di invarianza alla traslazione consegue che, se $h(n)$ è la risposta a $\delta(n)$, la risposta a $\delta(n-k)$ è semplicemente $h(n-k)$. La (1.6) diventa quindi

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \quad (1.7)$$

Ogni sistema lineare invariante alla traslazione è quindi completamente caratterizzato dalla risposta al campione unitario $h(n)$.

L'espressione (1.7) è comunemente chiamata *somma di convoluzione*. Se $y(n)$ è una sequenza i cui valori sono in relazione coi valori di due sequenze $h(n)$ e $x(n)$ secondo la (1.7), diciamo che $y(n)$ è la *convoluzione* di $x(n)$ con $h(n)$ e indichiamo questo con la notazione

$$y(n) = x(n) * h(n)$$

Con una sostituzione di variabili nella (1.7), otteniamo l'espressione alternativa

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = h(n) * x(n) \quad (1.8)$$

Quindi l'ordine secondo cui avviene la convoluzione delle due sequenze non è importante, e ne consegue che l'uscita del sistema è la stessa se i ruoli dell'ingresso e della risposta al campione unitario sono scambiati. In altri termini, un sistema lineare invariante alla traslazione con ingresso $x(n)$ e risposta all'impulso $h(n)$ avrà la stessa uscita di un sistema lineare invariante alla traslazione con ingresso $h(n)$ e risposta al campione unitario $x(n)$.

Due sistemi lineari invarianti alla traslazione in cascata formano un

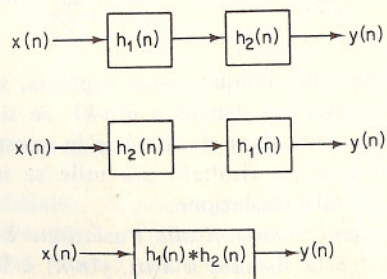


Fig. 1.5 Tre sistemi lineari invarianti alla traslazione con risposte al campione unitario identiche.

sistema lineare invariante alla traslazione con risposta al campione unitario che è la convoluzione delle due risposte. Dal momento che l'ordine in cui si esegue la convoluzione tra due sequenze non è importante, si ha che la risposta all'impulso di una combinazione a cascata di sistemi lineari invarianti alla traslazione è indipendente dall'ordine in cui essi sono posti in cascata. Questa proprietà è riassunta in fig. 1.5, in cui tutti e tre i sistemi rappresentati hanno la stessa risposta al campione unitario. Dalle (1.7) o (1.8) segue anche che due sistemi lineari invarianti alla traslazione in parallelo sono equivalenti a un unico sistema la cui risposta al campione unitario è la somma delle singole risposte. Ciò è rappresentato in fig. 1.6.

Sebbene l'espressione della somma di convoluzione sia analoga all'integrale di convoluzione della teoria dei sistemi lineari a tempo continuo, occorre sottolineare che la somma di convoluzione non deve essere considerata un'approssimazione dell'integrale di convoluzione. Vedremo che, mentre per i sistemi analogici l'integrale di convoluzione ha un'importanza quasi esclusivamente teorica, per i sistemi a tempo discreto la somma di convoluzione ha anche un'importanza pratica, in quanto può essere usata per realizzarli. Perciò è importante approfondire la comprensione delle proprietà della somma di convoluzione ed acquisire una certa agilità nell'usarla in calcoli effettivi.

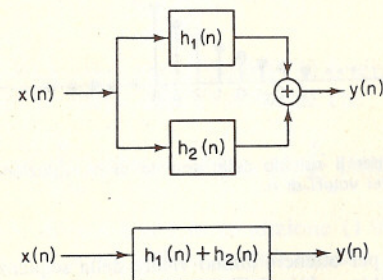


Fig. 1.6 Combinazione in parallelo di sistemi lineari invarianti alla traslazione e sistema equivalente.

ESEMPIO. Si consideri un sistema con risposta all'impulso

$$h(n) = \begin{cases} a^n, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

o, in maniera equivalente, $h(n) = a^n u(n)$. Per trovare la risposta ad un ingresso

$$x(n) = u(n) - u(n-N)$$

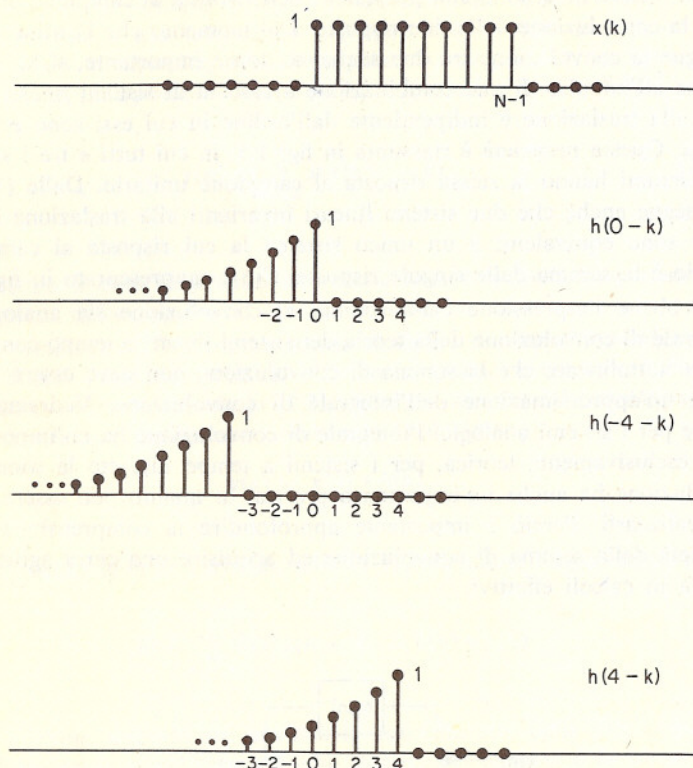


Fig. 1.7 Sequenze usate per il calcolo della somma di convoluzione, con $h(n-k)$ rappresentata per diversi valori di n .

notiamo dalla (1.7) che per ottenere l' n -mo valore della sequenza di uscita dobbiamo formare il prodotto $x(k)h(n-k)$ e sommare i valori della sequenza risultante. Le due sequenze componenti sono mostrate in fig. 1.7 in funzione di k , con $h(n-k)$ rappresentata per diversi valori di n . Come vediamo in fig. 1.7, $h(n-k)$ e $x(k)$ non hanno, per $n < 0$, campioni non nulli che si sovrappongono, e di conseguenza è $y(n) = 0$ per $n < 0$. Per n maggiore o eguale a zero ma minore di N , $h(n-k)$ e $x(k)$ hanno valori non nulli che si sovrappongono da $k = 0$ a $k = n$; perciò per $0 \leq n < N$,

$$y(n) = \sum_{k=0}^n a^{n-k} = a^n \frac{1 - a^{-(n+1)}}{1 - a^{-1}}, \quad 0 \leq n < N$$

Per $N-1 \leq n$ i valori non nulli che si sovrappongono si estendono da $k = 0$ a $k = N-1$, e perciò

$$y(n) = \sum_{k=0}^{N-1} a^{n-k} = a^n \frac{1 - a^{-N}}{1 - a^{-1}}, \quad N \leq n$$

La risposta $y(n)$ è rappresentata in fig. 1.8.

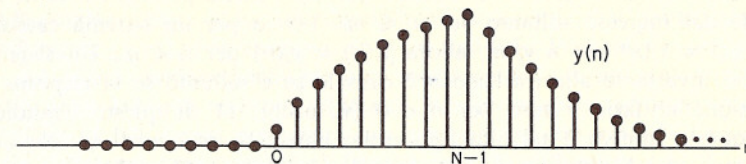


Fig. 1.8 Risposta all'ingresso $u(n) - u(n-N)$ del sistema con risposta al campione unitario $h(n) = a^n u(n)$.

1.3 STABILITÀ E CAUSALITÀ

Abbiamo visto che i vincoli di linearità e di invarianza nel tempo definiscono una classe di sistemi che sono rappresentati dalla somma di convoluzione. Gli ulteriori vincoli di stabilità e causalità definiscono una classe più ristretta di sistemi lineari invarianti nel tempo di importanza pratica.

BIBO Definiamo *sistema stabile* un sistema per cui ogni ingresso limitato provoca un'uscita limitata. I sistemi lineari invarianti alla traslazione sono stabili se e solo se

$$S \triangleq \sum_{k=-\infty}^{\infty} |h(k)| < \infty \quad (1.9)$$

Ciò può essere dimostrato come segue. Se la (1.9) è vera e x è limitata, cioè $|x(n)| < M$, per ogni n , allora dalla (1.8) segue che

$$|y(n)| = \left| \sum_{k=-\infty}^{\infty} h(k)x(n-k) \right| \leq M \sum_{k=-\infty}^{\infty} |h(k)| < \infty$$

Perciò y è limitata. Si dimostra che la condizione (1.9) è anche necessaria facendo vedere che se $S = \infty$, allora si può trovare un ingresso limitato che dà luogo ad un'uscita non limitata. Tale è la sequenza con valori

$$x(n) = \begin{cases} \frac{h^*(-n)}{|h(-n)|} & h(n) \neq 0 \\ 0, & h(n) = 0 \end{cases}$$

dove $h^*(n)$ è il complesso coniugato di $h(n)$. Chiaramente $x(n)$ è limitata. Il valore dell'uscita per $n = 0$ è

$$y(0) = \sum_{k=-\infty}^{\infty} x(-k)h(k) = \sum_{k=-\infty}^{\infty} \frac{|h(k)|^2}{|h(k)|} = S$$

Perciò se $S = \infty$, la sequenza di uscita è non limitata.

Un sistema causale è un sistema per cui l'uscita per ogni $n = n_0$ dipende dall'ingresso soltanto per $n \leq n_0$. Perciò per un sistema causale, se $x_1(n) = x_2(n)$ per $n < n_0$, allora $y_1(n) = y_2(n)$ per $n < n_0$. Un sistema lineare invariante alla traslazione è causale se e soltanto se la risposta al campione unitario è zero per $n < 0$ (v. probl. 11 di questo capitolo). Per questo motivo è talvolta opportuno chiamare *sequenza causale* una sequenza che è zero per $n < 0$, intendendo che essa potrebbe essere la risposta al campione unitario di un sistema causale.

Come esempio di stabilità e causalità, consideriamo il sistema lineare invariante alla traslazione con risposta al campione unitario $h(n) = a^n u(n)$; poiché tale risposta è zero per $n < 0$, il sistema è causale. Per determinare se si ha stabilità occorre calcolare la somma

$$S = \sum_{k=-\infty}^{\infty} |h(k)| = \sum_{k=0}^{\infty} |a|^k$$

Se $|a| < 1$, la serie geometrica ha per somma $S = 1/(1 - |a|)$. Tuttavia, se $|a| \geq 1$, la serie diverge e quindi il sistema è stabile soltanto per $|a| < 1$.

1.4 EQUAZIONI LINEARI ALLE DIFFERENZE A COEFFICIENTI COSTANTI

Esiste una sottoclasse di sistemi lineari invarianti alla traslazione che gioca un ruolo importante in molte applicazioni. Questa sottoclasse è costituita da quei sistemi per i quali l'ingresso $x(n)$ e l'uscita $y(n)$ soddisfano un'equazione alle differenze lineare a coefficienti costanti di ordine N di forma

$$\sum_{k=0}^N a_k y(n-k) = \sum_{r=0}^M b_r x(n-r) \quad (1.10)$$

In generale, un sistema di questa classe non è necessariamente causale. Per esempio, consideriamo l'equazione alle differenze del primo ordine

$$y(n) - ay(n-1) = x(n) \quad (1.11)$$

Si verifica facilmente per sostituzione diretta che se $x(n) = \delta(n)$, la (1.11) è soddisfatta sia da $y(n) = a^n u(n)$ che da $y(n) = a^n u(-n-1)$. La prima soluzione corrisponde a un filtro causale che è stabile se $|a| < 1$. La seconda soluzione è non causale ed è stabile soltanto se $|a| > 1$.

Comunemente un'equazione alle differenze come l'eq. (1.10) è interpretata come caratterizzante un sistema causale; ciò varrà generalmente anche per noi a meno che non si stabilisca diversamente.

Senza informazioni aggiuntive, un'equazione alle differenze della forma (1.10) non specifica univocamente la relazione ingresso-uscita di un

sistema. Questa è una conseguenza del fatto che, come per le equazioni differenziali, esiste tutta una famiglia di soluzioni. Per l'equazione alle differenze (1.11), per esempio, se $y_1(n)$ soddisfa l'equazione per $x(n) = x_1(n)$, questo avviene anche per ogni soluzione della forma $y(n) = y_1(n) + ka^n$, dove k è una costante arbitraria. Più in generale, ad ogni soluzione dell'eq. (1.10) si può aggiungere una componente che soddisfa l'equazione alle differenze omogenea (cioè l'equazione alle differenze con il secondo membro uguale a zero) e la somma soddisferà ancora l'eq. (1.10).

Un sistema che soddisfa un'equazione alle differenze lineare a coefficienti costanti corrisponderà ad un sistema lineare invariante alla traslazione soltanto se la componente omogenea viene scelta in maniera adeguata. Ad esempio, se il sistema è causale dobbiamo specificare condizioni iniziali di riposo in maniera che se $x(n) = 0$ per $n < n_0$, si abbia $y(n) = 0$ per $n < n_0$. Di qui in poi, a meno di indicazione diversa, assumeremo che, se un sistema soddisfa un'equazione alle differenze lineare a coefficienti costanti, soddisfa anche le condizioni necessarie per essere un sistema invariante alla traslazione.

Se supponiamo che il sistema sia causale, un'equazione lineare alle differenze fornisce una relazione esplicita tra l'ingresso e l'uscita. Ciò si può vedere riscrivendo l'eq. (1.10) come segue

$$y(n) = - \sum_{k=1}^N \frac{a_k}{a_0} y(n-k) + \sum_{r=0}^M \frac{b_r}{a_0} x(n-r) \quad (1.12)$$

Perciò l' n -mo valore dell'uscita può essere calcolato usando l' n -mo valore dell'ingresso e gli N ed M valori precedenti, rispettivamente dell'uscita e dell'ingresso. Come nel caso della somma di convoluzione, l'equazione alle differenze non rappresenta il sistema soltanto in linea teorica, ma può anche servire come realizzazione numerica del sistema.

ESEMPIO. Dall'equazione alle differenze del primo ordine (1.11) si deduce la formula ricorsiva

$$y(n) = ay(n-1) + x(n)$$

Per ottenere la risposta al campione unitario, si supponga $x(n) = \delta(n)$ e si assumano condizioni iniziali di riposo. Si ha allora

$$h(n) = 0, \quad n < 0$$

$$h(0) = ah(-1) + 1 = 1$$

$$h(1) = ah(0) = a$$

.

.

.

$$h(n) = ah(n-1) = a^n$$

Quindi

$$h(n) = a^n u(n)$$

per ottenere una soluzione diversa, si supponga $x(n) = \delta(n)$, ma si assuma ora che sia $y(n) = 0$ per $n > 0$. Dall'equazione alle differenze (1.11) possiamo scrivere la relazione ricorsiva

$$y(n-1) = \frac{1}{a}[y(n) - x(n)]$$

o

$$y(n) = \frac{1}{a}[y(n+1) - x(n+1)]$$

Si ha allora

$$h(n) = 0, \quad n > 0$$

$$h(0) = \frac{1}{a}[h(1) - x(1)] = 0$$

$$h(-1) = \frac{1}{a}[h(0) - x(0)] = -a^{-1}$$

$$h(-2) = \frac{1}{a}[h(-1) - x(-1)] = -a^{-2}$$

⋮

$$h(n) = \frac{1}{a} h(n+1) = -a^n$$

Quindi

$$h(n) = -a^n u(-n-1)$$

In generale, un sistema lineare invariante alla traslazione può avere una risposta al campione unitario di durata finita o infinita. A causa delle proprietà di alcune tecniche di elaborazione numerica, è utile distinguere tra queste due classi. Se la risposta all'impulso è di durata finita, parleremo di *sistema con risposta all'impulso finita (sistema FIR)* mentre se la risposta all'impulso è di durata infinita parleremo di *sistema con risposta all'impulso infinita (sistema IIR)*. Se $N = 0$ nell'eq. (1.10) cosicché

$$y(n) = \frac{1}{a_0} \left[\sum_{r=0}^M b_r x(n-r) \right]$$

allora essa corrisponde ad un sistema FIR. In effetti, il confronto con la (1.8) mostra che l'equazione alle differenze precedente è identica alla somma di convoluzione, e quindi ne segue direttamente che

$$h(n) = \begin{cases} \frac{b_n}{a_0}, & n = 0, 1, \dots, M \\ 0, & \text{altrove} \end{cases}$$

Un sistema FIR può sempre essere descritto da un'equazione alle differenze della forma (1.10) con $N = 0$. Invece per un sistema IIR N deve essere maggiore di zero.

1.5 RAPPRESENTAZIONE NEL DOMINIO DELLA FREQUENZA DI SISTEMI E SEGNALI A TEMPO DISCRETO

Nei paragrafi precedenti abbiamo introdotto alcuni dei concetti fondamentali della teoria dei segnali e dei sistemi a tempo discreto.

Per i sistemi lineari invarianti alla traslazione abbiamo visto che la rappresentazione della sequenza di ingresso come somma pesata di sequenze « campione unitario » ritardate [v. (1.4)] conduce ad una rappresentazione dell'uscita come somma pesata delle risposte al campione unitario ritardate. Come i segnali a tempo continuo, anche i segnali a tempo discreto possono essere rappresentati in modi differenti; in particolare le sequenze sinusoidali o esponenziali complesse giocano un ruolo particolarmente importante, proprio come per i sistemi e i segnali analogici.

Infatti i sistemi lineari invarianti alla traslazione hanno la proprietà fondamentale che la risposta a regime per un ingresso sinusoidale è sinusoidale alla stessa frequenza dell'ingresso, con ampiezza e fase determinate dal sistema. È questa proprietà dei sistemi lineari invarianti alla traslazione che rende così utile, nella teoria dei sistemi lineari, la rappresentazione dei segnali sotto forma di sinusoidi o di esponenziali complessi (rappresentazioni di Fourier).

Per vedere questo nel caso dei sistemi a tempo discreto, supponiamo che la sequenza d'ingresso sia $x(n) = e^{j\omega n}$ per $-\infty < n < \infty$, cioè un esponenziale complesso con pulsazione ω . Allora, utilizzando la (1.8), l'uscita è

$$\begin{aligned} y(n) &= \sum_{k=-\infty}^{\infty} h(k) e^{j\omega(n-k)} \\ &= e^{j\omega n} \sum_{k=-\infty}^{\infty} h(k) e^{-j\omega k} \end{aligned}$$

Se definiamo

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k) e^{-j\omega k} \quad (1.13)$$

si può scrivere

$$y(n) = H(e^{j\omega}) e^{j\omega n} \quad (1.14)$$

Dall'espressione (1.14) si vede che l'esponenziale complesso subisce semplicemente una variazione di ampiezza (complessa) descritta da $H(e^{j\omega})$ che è una funzione della frequenza. La quantità $H(e^{j\omega})$ è detta *risposta in frequenza* (o funzione di trasferimento) del sistema la cui risposta al campione unitario è $h(n)$. In generale $H(e^{j\omega})$ è complessa e può essere espressa in termini di parte reale e parte immaginaria come

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega})$$

o in termini di modulo e fase come

$$H(e^{j\omega}) = |H(e^{j\omega})| e^{j \arg[H(e^{j\omega})]}$$

Qualche volta sarà conveniente far riferimento al *ritardo di gruppo* piuttosto che alla fase. Il ritardo di gruppo è definito come la derivata prima rispetto a ω della fase, cambiata di segno.

Siccome una sinusoide può essere espressa come combinazione lineare di esponenziali complessi, la risposta in frequenza esprime anche la risposta all'ingresso sinusoidale. Più precisamente, si consideri

$$x(n) = A \cos(\omega_0 n + \phi) = \frac{A}{2} e^{j\phi} e^{j\omega_0 n} + \frac{A}{2} e^{-j\phi} e^{-j\omega_0 n}$$

Dalla (1.14) segue che la risposta a $(A/2)e^{j\phi}e^{j\omega_0 n}$ è

$$y_1(n) = H(e^{j\omega_0}) \frac{A}{2} e^{j\phi} e^{j\omega_0 n}$$

Se $h(n)$ è reale, dalla (1.7) si ricava che la risposta a $(A/2)e^{-j\phi}e^{-j\omega_0 n}$ è la complessa coniugata della risposta a $(A/2)e^{j\phi}e^{j\omega_0 n}$. Perciò la risposta complessiva è

$$y(n) = \frac{A}{2} [H(e^{j\omega_0}) e^{j\phi} e^{j\omega_0 n} + H(e^{-j\omega_0}) e^{-j\phi} e^{-j\omega_0 n}]$$

oppure

$$y(n) = A |H(e^{j\omega_0})| \cos(\omega_0 n + \phi + \theta)$$

dove $\theta = \arg[H(e^{j\omega_0})]$ indica la risposta di fase del sistema alla frequenza ω_0 .

$H(e^{j\omega})$ è una funzione continua di ω . Inoltre è funzione periodica di ω con periodo 2π . Questa proprietà deriva direttamente dalla (1.13) dato che $e^{j(\omega+2\pi)k} = e^{j\omega k}$. Il fatto che la risposta in frequenza abbia lo stesso valore per ω e per $\omega + 2\pi$ significa semplicemente che il sistema risponde in identica maniera agli esponenziali complessi di queste due frequenze. Siccome queste due sequenze esponenziali non sono distinguibili, tale comportamento è del tutto ragionevole.

ESEMPIO. Come esempio di calcolo della risposta in frequenza consideriamo un sistema con risposta al campione unitario

$$h(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (1.15)$$

descritta in fig. 1.9. La funzione di trasferimento è

$$\begin{aligned} H(e^{j\omega}) &= \sum_{n=0}^{N-1} e^{-j\omega n} = \frac{1 - e^{-j\omega N}}{1 - e^{-j\omega}} \\ &= \frac{\sin(\omega N/2)}{\sin(\omega/2)} e^{-j(N-1)\omega/2} \end{aligned} \quad (1.16)$$

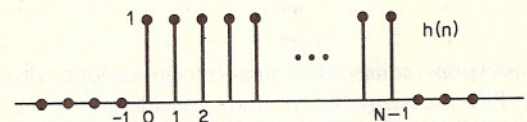


Fig. 1.9 Risposta all'impulso del sistema di cui si vuol calcolare la risposta in frequenza.

Il modulo e la fase di $H(e^{j\omega})$ sono riportati in fig. 1.10 per il caso di $N = 5$.

Siccome $H(e^{j\omega})$ è funzione periodica di ω , essa può essere rappresentata da una serie di Fourier. In effetti, la (1.13) esprime $H(e^{j\omega})$ nella forma di una serie di Fourier in cui i coefficienti di Fourier corri-

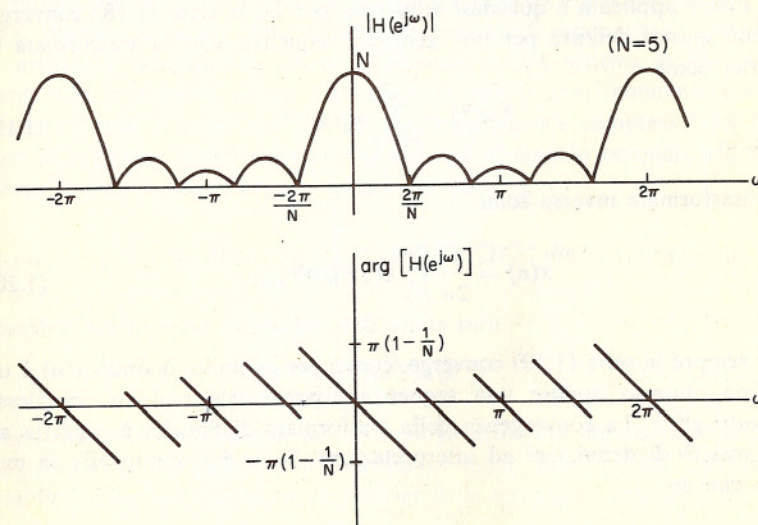


Fig. 1.10 Modulo e fase della risposta in frequenza del sistema la cui risposta al campione unitario è rappresentata in fig. 1.9.

spondono ai valori della risposta al campione unitario $h(n)$. Da questa osservazione deriva la possibilità di calcolare $h(n)$ da $H(e^{j\omega})$ per mezzo

della relazione usata per ottenere i coefficienti di Fourier di una funzione periodica [1-3], cioè

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{j\omega}) e^{j\omega n} d\omega \quad (1.17)$$

dove

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n) e^{-j\omega n} \quad (1.18)$$

Queste espressioni consentono una interpretazione alternativa della sequenza $h(n)$. Più precisamente, è utile considerare la (1.17) come la rappresentazione della sequenza $h(n)$ sotto forma di sovrapposizione (integrale) di segnali esponenziali, le cui ampiezze complesse sono fornite dalla (1.18). Perciò le (1.17) e (1.18) sono una coppia di trasformate di Fourier per la sequenza $h(n)$, dove la (1.18) gioca il ruolo di trasformata diretta (o analisi) della sequenza $h(n)$ e la (1.17) è la trasformata di Fourier inversa (sintesi). Tale rappresentazione esiste se la serie (1.18) converge.

La rappresentazione di una sequenza per mezzo della trasformata (1.18) non è limitata alla risposta al campione unitario di un sistema, ma può essere applicata a qualsiasi sequenza, purché la serie (1.18) converga. Si può quindi definire per una generica sequenza $x(n)$ la trasformata di Fourier come

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} \quad (1.19)$$

e la trasformata inversa come

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega \quad (1.20)$$

Non sempre la serie (1.19) converge, come, per esempio, quando $x(n)$ è un gradino unitario oppure una sequenza esponenziale reale o complessa per tutti gli n . La convergenza della trasformata di Fourier è soggetta ad una varietà di definizioni ed interpretazioni. Se $x(n)$ è sommabile in modulo, cioè se

$$\sum_{n=-\infty}^{\infty} |x(n)| < \infty$$

la serie si dice assolutamente convergente e converge uniformemente ad una funzione continua di ω . Di conseguenza la risposta in frequenza di un sistema stabile convergerà sempre. Se una sequenza è sommabile in mo-

dulo, essa avrà anche energia finita, cioè $\sum_{n=-\infty}^{\infty} |x(n)|^2$ sarà finita. Questo deriva direttamente dal fatto che

$$\sum |x(n)|^2 \leq [\sum |x(n)|]^2$$

Tuttavia non è vero che ogni sequenza con energia finita sia sommabile in modulo. Un esempio di sequenza che ha energia finita, ma che non è sommabile in modulo è dato dalla sequenza

$$x(n) = \frac{\sin \omega_0 n}{\pi n}$$

Se una sequenza non è sommabile in modulo, ma ha energia finita, si può usare un tipo di convergenza basato sul criterio dell'errore quadratico medio. Le oscillazioni di Gibbs che sono presenti in corrispondenza delle discontinuità [1-3] hanno importanza pratica nel progetto di filtri e saranno discusse in un capitolo successivo.

Il fatto che le sequenze possano essere rappresentate come sovrapposizione di esponenziali complessi è molto importante nell'analisi dei sistemi lineari invarianti alla traslazione, in virtù del principio di sovrapposizione e del fatto che la risposta di tali sistemi ad esponenziali complessi è completamente determinata dalla risposta in frequenza, $H(e^{j\omega})$. Se si intende la (1.20) come la sovrapposizione di esponenziali complessi di ampiezza infinitesima, allora la risposta di un sistema lineare invariante alla traslazione ad un ingresso $x(n)$ sarà la corrispondente sovrapposizione delle risposte relative ad ogni esponenziale complesso che compare in ingresso; quindi, siccome la risposta al singolo esponenziale complesso si ottiene moltiplicando questo per $H(e^{j\omega})$,

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{j\omega}) X(e^{j\omega}) e^{j\omega n} d\omega$$

Perciò la trasformata di Fourier dell'uscita sarà

$$Y(e^{j\omega}) = H(e^{j\omega}) X(e^{j\omega}) \quad (1.21)$$

Questo risultato corrisponde ad un'analoga proprietà dei sistemi lineari a tempo continuo e può essere ricavato in modo più rigoroso calcolando semplicemente la trasformata di Fourier della somma di convoluzione

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k) x(k)$$

Sebbene questo approccio più formale fornisca una giustificazione rigorosa della (1.21) (v. probl. 17 di questo capitolo), la discussione precedente ha lo scopo di mettere in evidenza che la (1.21) è una diretta conseguenza delle particolari proprietà dei sistemi lineari invarianti alla traslazione.

Per illustrare i risultati di cui si sta discutendo, consideriamo il seguente esempio.

ESEMPIO. Filtro passa-basso ideale. Il filtro ideale passa-basso a tempo discreto ha risposta in frequenza $H(e^{j\omega})$, rappresentata in fig. 1.11; vale a dire, per $-\pi < \omega < \pi$,

$$H(e^{j\omega}) = \begin{cases} 1, & |\omega| \leq \omega_{co} \\ 0, & \omega_{co} < |\omega| \leq \pi \end{cases}$$

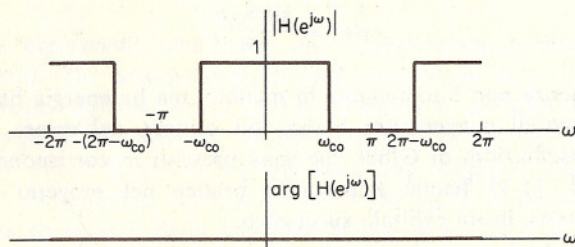


Fig. 1.11 Risposta in frequenza di un filtro ideale passa-basso a tempo discreto.

Siccome $H(e^{j\omega})$ è periodica, questa relazione definisce la risposta in frequenza per tutti i valori di ω . Tale sistema chiaramente elimina tutte le componenti di frequenza in ingresso comprese nel campo $\omega_{co} < |\omega| \leq \pi$. La risposta all'impulso $h(n)$ ricavata dalla (1.17) è

$$h(n) = \frac{1}{2\pi} \int_{-\omega_{co}}^{\omega_{co}} e^{j\omega n} d\omega = \frac{\sin \omega_{co} n}{\pi n}$$

ed è schematizzata in fig. 1.12 per il caso $\omega_{co} = \pi/2$.

Il filtro passa-basso ideale è l'esempio di un sistema che può essere descritto molto efficacemente nel dominio della frequenza. Come si vede facilmente, il sistema elimina completamente tutte le frequenze in ingresso al di sopra della frequenza di taglio ω_{co} . Chiaramente il filtro ideale passa-basso non è causale; inoltre si può dimostrare la sua instabilità in senso stretto secondo la definizione del par. 1.3.

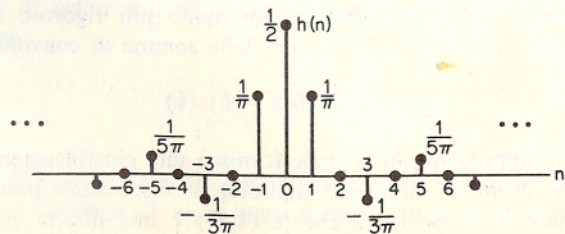


Fig. 1.12 Risposta all'impulso del filtro ideale passa-basso con frequenza di taglio $\omega_{co} = \pi/2$.

Ciò nonostante, tale sistema è estremamente importante da un punto di vista concettuale, e nel cap. 5 si dedicheranno molti sforzi alla ricerca di metodi per progettare sistemi che possano approssimare il comportamento del filtro ideale passa-basso.

1.6 ALCUNE PROPRIETÀ DI SIMMETRIA DELLA TRASFORMATTA DI FOURIER

Esistono delle proprietà di simmetria della trasformata di Fourier che sono spesso molto utili. In questo paragrafo vengono presentate alcune di queste proprietà, mentre le dimostrazioni sono affrontate nei problemi 25-27 di questo capitolo.

Si definisce *coniugata simmetrica* una sequenza $x_e(n)$ per cui $x_e(n) = x_e^*(-n)$ e si definisce *coniugata antisimmetrica* una sequenza $x_o(n)$ per la quale $x_o(n) = -x_o^*(-n)$, dove $*$ significa coniugato. Una sequenza arbitraria $x(n)$ può essere sempre espressa come la somma di una sequenza coniugata simmetrica e di una coniugata antisimmetrica:

$$x(n) = x_e(n) + x_o(n) \quad (1.22a)$$

dove

$$x_e(n) = \frac{1}{2}[x(n) + x^*(-n)] \quad (1.22b)$$

e

$$x_o(n) = \frac{1}{2}[x(n) - x^*(-n)] \quad (1.22c)$$

Una sequenza reale che sia coniugata simmetrica, cioè tale che $x_e(n) = x_e(-n)$, viene detta generalmente *sequenza pari*, ed una sequenza reale che sia coniugata antisimmetrica, cioè tale che $x_o(n) = -x_o(-n)$, è chiamata generalmente *sequenza dispari*.

Una trasformata di Fourier $X(e^{j\omega})$ può essere scomposta nella somma di una funzione coniugata simmetrica e di una coniugata antisimmetrica secondo la

$$X(e^{j\omega}) = X_e(e^{j\omega}) + X_o(e^{j\omega}) \quad (1.23a)$$

dove

$$X_e(e^{j\omega}) = \frac{1}{2}[X(e^{j\omega}) + X^*(e^{-j\omega})] \quad (1.23b)$$

e

$$X_o(e^{j\omega}) = \frac{1}{2}[X(e^{j\omega}) - X^*(e^{-j\omega})] \quad (1.23c)$$

dove $X_e(e^{j\omega})$ è coniugata simmetrica e $X_o(e^{j\omega})$ è coniugata antisimmetrica, cioè

$$X_e(e^{j\omega}) = X_e^*(e^{-j\omega})$$

e

$$X_o(e^{j\omega}) = -X_o^*(e^{-j\omega})$$

Come nel caso delle sequenze, se una funzione reale è coniugata simmetrica è detta generalmente *funzione pari* e se coniugata antisimmetrica *funzione dispari*.

Consideriamo dapprima una sequenza complessa generica $x(n)$ con trasformata di Fourier $X(e^{j\omega})$. Si può dimostrare (v. probl. 25) che la trasformata di Fourier di $x^*(n)$ è $X^*(e^{-j\omega})$ e che la trasformata di Fourier di $x^*(-n)$ è $X^*(e^{j\omega})$. Da questo, e dal fatto che la trasformata di Fourier della somma di due sequenze è la somma delle trasformate di Fourier, segue che la trasformata di Fourier di $\frac{1}{2}[x(n) + x^*(n)]$ ovvero di $\text{Re}[x(n)]$, è $\frac{1}{2}[X(e^{j\omega}) + X^*(e^{-j\omega})]$ cioè la parte coniugata simmetrica di $X(e^{j\omega})$. Analogamente, $\frac{1}{2}[x(n) - x^*(n)]$ ovvero $j\text{Im}[x(n)]$, ha una trasformata di Fourier che è la componente coniugata antisimmetrica $X_o(e^{j\omega})$. Considerando ora le trasformate di Fourier di $x_e(n)$ e di $x_o(n)$, le componenti coniugata simmetrica e coniugata antisimmetrica di $x(n)$, risulta che la trasformata di Fourier di $x_e(n)$ è $\text{Re}[X(e^{j\omega})]$ e la trasformata di Fourier di $x_o(n)$ è $j\text{Im}[X(e^{j\omega})]$.

Se $x(n)$ è una sequenza reale, queste proprietà di simmetria diventano particolarmente semplici e utili. Nella fattispecie, per una sequenza reale, la trasformata di Fourier è coniugata simmetrica, cioè $X(e^{j\omega}) = X^*(e^{-j\omega})$. Esprimendo $X(e^{j\omega})$ in termini della sua parte reale e immaginaria come

$$X(e^{j\omega}) = \text{Re}[X(e^{j\omega})] + j\text{Im}[X(e^{j\omega})]$$

segue che

$$\text{Re}[X(e^{j\omega})] = \text{Re}[X(e^{-j\omega})]$$

e

$$\text{Im}[X(e^{j\omega})] = -\text{Im}[X(e^{-j\omega})]$$

cioè la parte reale della trasformata di Fourier è una funzione pari e la parte immaginaria è una funzione dispari. In maniera analoga, esprimendo $X(e^{j\omega})$ in coordinate polari come

$$X(e^{j\omega}) = |X(e^{j\omega})| e^{j\arg[X(e^{j\omega})]}$$

segue che per una sequenza reale $x(n)$ il modulo della trasformata di Fourier $|X(e^{j\omega})|$ è una funzione pari di ω e la fase, data da $\arg[X(e^{j\omega})]$, è una funzione dispari di ω . Ancora per una sequenza reale, la parte pari di $x(n)$ ha per trasformata $\text{Re}[X(e^{j\omega})]$ e la parte dispari di $x(n)$ ha per trasformata $j\text{Im}[X(e^{j\omega})]$. Tutte le proprietà di simmetria esposte sono riassunte nella tabella 1.1.

Tab. 1.1

Sequenza	Trasformata di Fourier
$x(n)$	$X(e^{j\omega})$
$x^*(n)$	$X^*(e^{-j\omega})$
$x^*(-n)$	$X^*(e^{j\omega})$
$\text{Re}[x(n)]$	$X_e(e^{j\omega})$ [parte coniugata simmetrica di $X(e^{j\omega})$]
$j\text{Im}[x(n)]$	$X_o(e^{j\omega})$ [parte coniugata antisimmetr. di $X(e^{j\omega})$]
$x_e(n)$ [Parte coniugata simmetr. di $x(n)$]	$\text{Re}[X(e^{j\omega})]$
$x_o(n)$ [Parte coniugata antisimmetr. di $x(n)$]	$j\text{Im}[X(e^{j\omega})]$

Le seguenti proprietà valgono solo per $x(n)$ reale:

$x(n)$ reale qualsiasi	$X(e^{j\omega}) = X^*(e^{-j\omega})$ (La trasformata di Fourier è coniugata simmetrica)
	$\text{Re}[X(e^{j\omega})] = \text{Re}[X(e^{-j\omega})]$ (la parte reale è pari)
	$\text{Im}[X(e^{j\omega})] = -\text{Im}[X(e^{-j\omega})]$ (la parte immaginaria è dispari)
	$ X(e^{j\omega}) = X(e^{-j\omega}) $ (il modulo è pari)
	$\arg[X(e^{j\omega})] = -\arg[X(e^{-j\omega})]$ (la fase è dispari)
$x_e(n)$ [parte pari di $x(n)$]	$\text{Re}[X(e^{j\omega})]$
$x_o(n)$ [parte dispari di $x(n)$]	$j\text{Im}[X(e^{j\omega})]$

1.7 CAMPIONAMENTO DI SEGNALI A TEMPO CONTINUO

Nei paragrafi precedenti di questo capitolo abbiamo evitato di porre in relazione segnali e sistemi a tempo discreto con quelli a tempo continuo, eccetto quando abbiamo sottolineato certe analogie tra concetti teorici importanti. Capita spesso, tuttavia, che i segnali a tempo discreto siano derivati da quelli a tempo continuo mediante un campionamento periodico: di conseguenza è importante capire come sia legata al segnale originale la sequenza così ottenuta.

Si consideri un segnale analogico $x_a(t)$ che ha la rappresentazione di Fourier [1-3]

$$x_a(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X_a(j\Omega) e^{j\Omega t} d\Omega \quad (1.24a)$$

$$X_a(j\Omega) = \int_{-\infty}^{\infty} x_a(t) e^{-j\Omega t} dt \quad (1.24b)$$

La sequenza $x(n)$ con valori $x(n) = x_a(nT)$ si dice essere ottenuta da $x_a(t)$ mediante campionamento periodico, e T è chiamato *periodo di campionamento*. L'inverso di T viene detto *frequenza di campionamento* o *velocità di campionamento*. Per trovare in che senso $x(n)$ rappresenta il segnale originario $x_a(t)$, conviene mettere in relazione $X_a(j\Omega)$, cioè la trasformata di Fourier a tempo continuo di $x_a(t)$, con $X(e^{j\omega})$, cioè la trasformata di Fourier a tempo discreto della sequenza $x(n)$. In base alla (1.24a) notiamo che

$$x(n) = x_a(nT) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_a(j\Omega) e^{j\Omega nT} d\Omega \quad (1.25)$$

Ma, a partire dalla trasformata di Fourier a tempo discreto, possiamo anche scrivere

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega \quad (1.26)$$

Per collegare le (1.25) e (1.26) conviene esprimere la (1.25) come una somma di integrali sopra intervalli di durata $2\pi/T$, come in

$$x(n) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \int_{(2r-1)\pi/T}^{(2r+1)\pi/T} X_a(j\Omega) e^{j\Omega nT} d\Omega$$

Con un cambiamento di variabile ogni termine della sommatoria può essere scritto come un integrale nell'intervallo $(-\pi/T, +\pi/T)$ ottenendo

$$x(n) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \int_{-\pi/T}^{\pi/T} X_a\left(j\Omega + j\frac{2\pi r}{T}\right) e^{j\Omega nT} e^{j2\pi r n} d\Omega$$

Se ora scambiamo l'ordine di integrale e sommatoria e teniamo conto che $e^{j2\pi r n} = 1$ per tutti i valori interi di r ed n , otteniamo

$$x(n) = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} \left[\sum_{r=-\infty}^{\infty} X_a\left(j\Omega + j\frac{2\pi r}{T}\right) \right] e^{j\Omega nT} d\Omega \quad (1.27)$$

Con la sostituzione $\Omega = \omega/T$, la (1.27) diventa

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{1}{T} \sum_{r=-\infty}^{\infty} X_a\left(\frac{j\omega}{T} + j\frac{2\pi r}{T}\right) \right] e^{j\omega n} d\omega$$

che è identica nella forma alla (1.26). Perciò possiamo dedurre l'identità

$$X(e^{j\omega}) = \frac{1}{T} \sum_{r=-\infty}^{\infty} X_a\left(\frac{j\omega}{T} + j\frac{2\pi r}{T}\right) \quad (1.28)$$

Se invece si usa come variabile la frequenza analogica Ω , questa espressione diventa

$$X(e^{j\Omega T}) = \frac{1}{T} \sum_{r=-\infty}^{\infty} X_a\left(j\Omega + j\frac{2\pi r}{T}\right) \quad (1.29)$$

Le (1.28) e (1.29) esprimono chiaramente la relazione esistente fra la trasformata di Fourier a tempo continuo e la trasformata di Fourier di una

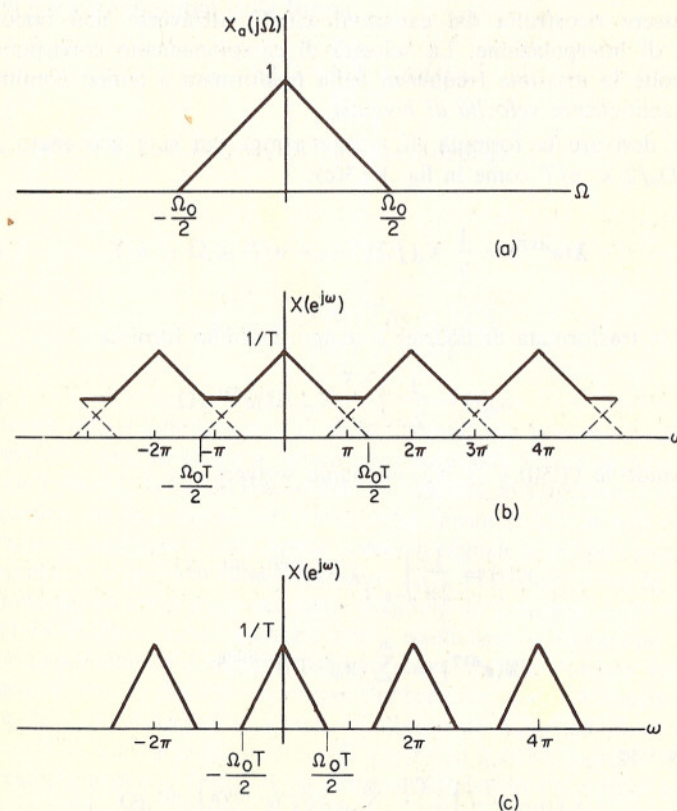


Fig. 1.13 (a) Trasformata di Fourier di un segnale a tempo continuo. (b) Trasformata di Fourier del segnale a tempo discreto ottenuto con campionamento periodico. Il periodo di campionamento è così grande da provocare sovrapposizioni delle ripetizioni periodiche della trasformata a tempo continuo. (c) Come (b), con la differenza che il periodo di campionamento è abbastanza piccolo da non provocare sovrapposizioni delle ripetizioni periodiche della trasformata a tempo continuo.

sequenza derivata mediante campionamento. Per esempio, se $X_a(j\Omega)$ è la funzione rappresentata in fig. 1.13 (a), allora $X(e^{j\omega})$ sarà la funzione rappresentata in fig. 1.13(b) se $\Omega_0/2 > \pi/T$. La fig. 1.13(b) mette in evidenza che, se il periodo di campionamento è troppo lungo, le repliche traslate di $X_a(j\omega/T)$ si sovrappongono. In tal caso le alte frequenze di $X_a(j\Omega)$ si riflettono nelle basse frequenze di $X(e^{j\omega})$. Questo fenomeno, dove in effetti una componente ad alta frequenza in $X_a(j\Omega)$ si trasforma in una componente di più bassa frequenza, viene chiamato *aliasing*. Dalla fig. 1.13(c) risulta chiaro che, se $\Omega_0/2 < \pi/T$, cioè, se campioniamo ad una frequenza che è almeno due volte la più alta frequenza di $X_a(j\Omega)$, allora $X(e^{j\omega})$ risulta identica a $X_a(\omega/T)$ nell'intervallo $-\pi \leq \omega \leq +\pi$. In tal caso è ragionevole attendersi che $x_a(t)$ possa essere ricostruita dai campioni $x_a(nT)$ attraverso una opportuna formula di interpolazione. La velocità di campionamento corrispondente a due volte la massima frequenza della trasformata a tempo continuo si chiama solitamente *velocità di Nyquist*.

Per derivare la formula di interpolazione cui si è accennato, assumiamo $\Omega_0/2 < \pi/T$ come in fig. 1.13(c). Allora

$$X(e^{j\Omega T}) = \frac{1}{T} X_a(j\Omega), \quad -\pi/T \leq \Omega \leq \pi/T \quad (1.30)$$

Inoltre, la trasformata di Fourier a tempo continuo fornisce

$$x_a(t) = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} X_a(j\Omega) e^{j\Omega t} d\Omega \quad (1.31)$$

Combinando le (1.30) e (1.31), possiamo scrivere

$$x_a(t) = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} T X(e^{j\Omega T}) e^{j\Omega t} d\Omega$$

Poiché

$$X(e^{j\Omega T}) = \sum_{k=-\infty}^{\infty} x_a(kT) e^{-j\Omega T k}$$

ne segue che

$$x_a(t) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \left[\sum_{k=-\infty}^{\infty} x_a(kT) e^{-j\Omega T k} \right] e^{j\Omega t} d\Omega$$

oppure, scambiando l'ordine di sommatoria e di integrale,

$$x_a(t) = \sum_{k=-\infty}^{\infty} x_a(kT) \left[\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} e^{j\Omega(t-kT)} d\Omega \right]$$

che, valutando l'integrale, diventa

$$x_a(t) = \sum_{k=-\infty}^{\infty} x_a(kT) \frac{\sin[(\pi/T)(t-kT)]}{(\pi/T)(t-kT)} \quad (1.32)$$

L'espressione (1.32) fornisce una formula di interpolazione per ricostruire il segnale a tempo continuo $x_a(t)$ a partire dai suoi campioni. La rappresentazione di un segnale a tempo continuo nella forma (1.32) vale solo per funzioni limitate in banda e con T scelto sufficientemente piccolo da non provocare *aliasing*.

L'espressione (1.32) può essere pensata come un'espansione del segnale a tempo continuo nella forma

$$x_a(t) = \sum_{k=-\infty}^{\infty} c_k \phi_k(t) \quad (1.33)$$

dove i coefficienti c_k e le funzioni $\phi_k(t)$ sono dati da

$$c_k = x_a(kT) \quad (1.34a)$$

e

$$\phi_k(t) = \frac{\sin[(\pi/T)(t-kT)]}{(\pi/T)(t-kT)} \quad (1.34b)$$

In generale esistono molte classi di funzioni $\phi_k(t)$ che possono essere usate per esprimere una funzione a tempo continuo nella forma (1.33); queste classi includono le funzioni sinusoidali, le funzioni di Laguerre, i polinomi di Legendre. In ogni rappresentazione della forma (1.33) la sequenza dei coefficienti c_k può essere considerata come un segnale a tempo discreto che rappresenta il segnale a tempo continuo $x_a(t)$ [7]. Non tutte le rappresentazioni di questo tipo sono utili. Scegliere le funzioni $\phi_k(t)$ espresse dalla (1.34 b) dà il forte vantaggio che i coefficienti si ottengono semplicemente campionando il segnale a tempo continuo. Un ulteriore vantaggio di questa scelta è che essa conserva la convoluzione nel senso che se $y_a(t)$ è la convoluzione a tempo continuo di $x_a(t)$ ed $h_a(t)$, allora $y_a(nT)$ sarà la convoluzione discreta di $x_a(nT)$ ed $h_a(nT)$, purché il periodo di campionamento T sia scelto abbastanza piccolo da evitare *aliasing*, ovvero in modo che la rappresentazione (1.32) sia valida. Il vantaggio di una rappresentazione che conserva la convoluzione è che essa consente la simulazione o realizzazione di un sistema lineare a tempo continuo e tempo-invariante per mezzo di un sistema lineare a tempo discreto e invariante alla traslazione. Lo svantaggio della rappresentazione che usa le funzioni (1.34 b) è che essa può applicarsi solo alle funzioni limitate in banda.

Esistono molte altre scelte per le funzioni $\phi_i(t)$ della (1.33) tali da fornire una rappresentazione discreta di segnali a tempo continuo che conserva la convoluzione [8]. Una condizione necessaria e sufficiente a questo scopo verrà presa in considerazione nel probl. 32 di questo stesso capitolo.

1.8 SEQUENZE E SISTEMI BIDIMENSIONALI

Nel paragrafo precedente abbiamo concentrato l'attenzione sulla rappresentazione di segnali e sistemi monodimensionali. In questo paragrafo prenderemo in considerazione l'estensione di alcuni dei precedenti risultati al caso di segnali e sistemi bidimensionali. Molti problemi importanti di elaborazione dei segnali comportano la trattazione di segnali multidimensionali. Tutte le proprietà dei segnali e dei sistemi che sono state fino a qui ricavate in questo capitolo sono di facile estensione al caso multidimensionale. Nei capitoli che seguono, tuttavia, non succederà, in generale, che risultati validi per sequenze e sistemi monodimensionali siano estendibili al caso multidimensionale.

Una sequenza bidimensionale è una funzione di due variabili intere ed è spesso rappresentata graficamente come è mostrato in fig. 1.14. Come nel caso monodimensionale, è utile definire il campione unitario, il gradino unitario, la sequenza esponenziale e la sequenza sinusoidale. La sequenza campione unitario bidimensionale $\delta(m, n)$ è definita come la sequenza che è nulla ovunque fuorché nell'origine, cioè

$$\delta(m, n) = \begin{cases} 0, & m, n \neq 0 \\ 1, & m = n = 0 \end{cases}$$

La sequenza gradino unitario bidimensionale $u(m, n)$ è uguale ad uno

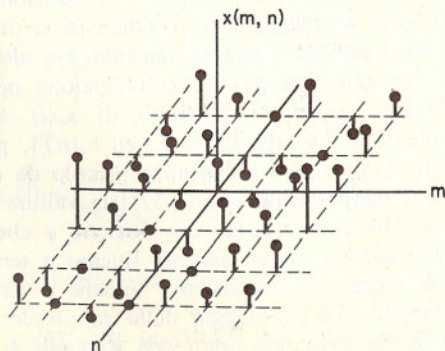


Fig. 1.14 Rappresentazione grafica di una sequenza bidimensionale.

nel primo quadrante del piano (m, n) ed è nulla altrove, cioè

$$u(m, n) = \begin{cases} 1, & m \geq 0, n \geq 0 \\ 0, & \text{altrove} \end{cases}$$

Una sequenza esponenziale bidimensionale ha la forma $a^m b^n$, e una sequenza sinusoidale bidimensionale ha la forma

$$A \cos(\omega_0 m + \phi) \cos(\omega_1 n + \theta).$$

Una sequenza separabile è una sequenza che può essere espressa come prodotto di sequenze monodimensionali, cioè $x(m, n)$ è separabile se può essere espressa nella forma

$$x(m, n) = x_1(m)x_2(n)$$

Le sequenze campione unitario, gradino unitario, esponenziale e sinusoidale sono tutte separabili. Un esempio di sequenza non separabile è la sequenza

$$x(m, n) = \cos(\omega_0 mn)$$

Come nel caso monodimensionale, una sequenza bidimensionale qualsiasi può essere espressa come combinazione lineare di campioni unitari traslati:

$$x(m, n) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} x(k, r) \delta(m - k, n - r)$$

Sulla base di questa equazione, un sistema lineare bidimensionale può essere descritto in termini delle sue risposte a campioni unitari traslati. Più precisamente, con $y(m, n) = T[x(m, n)]$, dove $T[\]$ è la trasformazione per un sistema lineare, si ha

$$\begin{aligned} y(m, n) &= T \left[\sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} x(k, r) \delta(m - k, n - r) \right] \\ &= \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} x(k, r) T[\delta(m - k, n - r)] \end{aligned}$$

Indicando con $h_{k,r}(m, n)$ la risposta del sistema a $\delta(m - k, n - r)$, si può scrivere

$$y(m, n) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} x(k, r) h_{k,r}(m, n) \quad (1.35)$$

Con la sola condizione di linearità imposta sul sistema, $h_{k,r}(m, n)$ dipenderà dalle quattro variabili k, r, m ed n . È tuttavia utile imporre l'ulteriore vincolo di invarianza alla traslazione. La classe dei sistemi bidimensionali

invarianti alla traslazione è caratterizzata dalla proprietà che, se $y(m, n)$ è la risposta ad $x(m, n)$, allora $y(m-k, n-r)$ è la risposta ad $x(m-k, n-r)$. Con questo vincolo sul sistema, se $h(m, n)$ è la risposta a $\delta(m, n)$, allora $h(m-k, n-r)$ è la risposta a $\delta(m-k, n-r)$. In questo caso, la (1.35) diventa

$$y(m, n) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} x(k, r) h(m-k, n-r) \quad (1.36)$$

L'espressione (1.36) è la *somma di convoluzione* per un sistema bidimensionale lineare invariante alla traslazione. Con un cambiamento di variabili nella (1.36) si ottiene l'espressione alternativa

$$y(m, n) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(k, r) x(m-k, n-r) \quad (1.37)$$

Perciò, come nel caso monodimensionale, l'operazione di convoluzione tra due sequenze è commutativa; cioè, non ha importanza l'ordine in cui si effettua la convoluzione. Ciò implica, tra l'altro, che la risposta all'impulso della cascata di due sistemi lineari invariati alla traslazione è indipendente dall'ordine in cui essi sono messi in cascata.

Un sistema stabile è un sistema per cui ogni ingresso limitato produce un'uscita limitata. Ricalcando il ragionamento del par. 1.3 (v. probl. 39 di questo capitolo), si può mostrare che un sistema bidimensionale lineare invariante alla traslazione è stabile se e solo se

$$S \triangleq \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} |h(k, r)| < \infty \quad (1.38)$$

Un sistema bidimensionale si dice causale se, quando due ingressi $x_1(m, n)$ ed $x_2(m, n)$ sono uguali per $(m < m_1, n < n_1)$, allora le corrispondenti uscite $y_1(m, n)$ ed $y_2(m, n)$ sono uguali per $(m < m_1, n < n_1)$. Per un sistema lineare invariante alla traslazione, la causalità implica che la risposta al campione unitario sia zero per $(m < 0, n < 0)$. Inversamente, se la risposta al campione unitario è zero per $(m < 0, n < 0)$, allora il sistema è causale.

Una sottoclasse importante di sistemi bidimensionali lineari invarianti alla traslazione è costituita da quei sistemi per cui l'ingresso e l'uscita soddisfano un'equazione alle differenze lineare a coefficienti costanti della forma

$$\sum_{k=0}^{M_1} \sum_{r=0}^{N_1} a_{kr} y(m-k, n-r) = \sum_{k=0}^{M_2} \sum_{r=0}^{N_2} b_{kr} x(m-k, n-r) \quad (1.39)$$

dove, per la linearità, imponiamo l'ulteriore condizione che, se $x(m, n) = 0$ per tutti gli m ed n , allora $y(m, n) = 0$ per tutti gli m ed n . Come con le equazioni alle differenze monodimensionali, l'equazione alle differenze bidimensionale (1.39) può corrispondere ad un sistema causale o non causale. Se supponiamo che il sistema sia causale, la (1.39) può essere espressa come una relazione ricorsiva, cioè

$$y(m, n) = \frac{1}{a_{00}} \left\{ \sum_{k=0}^{M_2} \sum_{r=0}^{N_2} b_{kr} x(m-k, n-r) - \sum_{\substack{k=0 \\ k, r \neq 0}}^{M_1} \sum_{r=0}^{N_1} a_{kr} y(m-k, n-r) \right\} \quad (1.40)$$

$k, r \neq 0$ contemporaneamente

Ad esempio, se $x(m, n)$ è zero per $m < 0, n < 0$, allora, in conseguenza della causalità, $y(m, n)$ è zero per $m < 0, n < 0$. Questo fornisce quindi un insieme di condizioni iniziali da usare nell'iterazione della (1.40).

Per sistemi bidimensionali lineari invarianti alla traslazione, la risposta ad un esponenziale complesso della forma $e^{j\omega_1 m} e^{j\omega_2 n}$ è un esponenziale complesso delle stesse frequenze complesse. Specificamente, se è

$$x(m, n) = e^{j\omega_1 m} e^{j\omega_2 n}$$

allora dalla somma di convoluzione si ha

$$y(m, n) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(k, r) e^{j\omega_1 m} e^{j\omega_2 n} e^{-j\omega_1 k} e^{-j\omega_2 r} = H(e^{j\omega_1}, e^{j\omega_2}) e^{j\omega_1 m} e^{j\omega_2 n}$$

dove

$$H(e^{j\omega_1}, e^{j\omega_2}) = \sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(k, r) e^{-j\omega_1 k} e^{-j\omega_2 r} \quad (1.41)$$

$H(e^{j\omega_1}, e^{j\omega_2})$ è la risposta in frequenza del sistema bidimensionale. Essa è una funzione continua di ω_1 ed ω_2 ed è una funzione periodica di ciascuna di queste variabili con periodo 2π . Si può mostrare (v. probl. 38 di questo capitolo) che se $h(m, n)$ è separabile, allora $H(e^{j\omega_1}, e^{j\omega_2})$ è

separabile, cioè può essere espressa nella forma

$$H(e^{j\omega_1}, e^{j\omega_2}) = H_1(e^{j\omega_1})H_2(e^{j\omega_2})$$

La sequenza $h(m, n)$ può essere ricostruita dalla risposta in frequenza per mezzo della relazione

$$h(m, n) = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} H(e^{j\omega_1}, e^{j\omega_2}) e^{j\omega_1 m} e^{j\omega_2 n} d\omega_1 d\omega_2 \quad (1.42)$$

Più in generale, definiamo la trasformata di Fourier bidimensionale di una sequenza $x(m, n)$ come

$$X(e^{j\omega_1}, e^{j\omega_2}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(m, n) e^{-j\omega_1 m} e^{-j\omega_2 n} \quad (1.43)$$

con la relazione inversa data da

$$x(m, n) = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} X(e^{j\omega_1}, e^{j\omega_2}) e^{j\omega_1 m} e^{j\omega_2 n} d\omega_1 d\omega_2 \quad (1.44)$$

Applicando la trasformata (1.43) alla somma di convoluzione, si vede che le trasformate di Fourier dell'ingresso e dell'uscita di un sistema bidimensionale lineare invariante alla traslazione sono legate dalla relazione

$$Y(e^{j\omega_1}, e^{j\omega_2}) = H(e^{j\omega_1}, e^{j\omega_2}) X(e^{j\omega_1}, e^{j\omega_2}) \quad (1.45)$$

SOMMARIO

In questo capitolo abbiamo considerato un certo numero di nozioni di base relative a segnali e sistemi a tempo discreto. In particolare, dopo aver preso in esame diverse sequenze elementari, abbiamo considerato la definizione e la rappresentazione di sistemi lineari invarianti alla traslazione in termini della somma di convoluzione, nonché alcune implicazioni delle proprietà di stabilità e causalità. Una sottoclasse importante di sistemi li-

neari invarianti alla traslazione è quella per cui l'ingresso e l'uscita soddisfano un'equazione alle differenze lineare a coefficienti costanti. È stata discussa la soluzione iterativa di tali equazioni e sono state definite le classi di sistemi FIR e IIR.

Un importante strumento di analisi e rappresentazione dei sistemi lineari invarianti alla traslazione è costituito dalla loro rappresentazione nel dominio della frequenza. È stata considerata la risposta di un sistema ad un ingresso esponenziale complesso, che conduce alla definizione della risposta in frequenza. In particolare, si è trovato che la risposta all'impulso e la risposta in frequenza sono legate tra loro come una coppia di trasformate di Fourier. Sono poi state sviluppate alcune proprietà delle trasformate di Fourier.

Sebbene il materiale di questo capitolo sia stato presentato senza esplicito riferimento ai segnali a tempo continuo, una classe importante di problemi di elaborazione numerica dei segnali ha origine dal campionamento dei segnali continui. Di conseguenza, nel par. 1.7 abbiamo considerato la relazione tra segnali a tempo continuo e sequenze ottenute con un campionamento periodico.

Il capitolo si è chiuso con una breve introduzione alle sequenze e ai sistemi bidimensionali.

BIBLIOGRAFIA

1. E. A. Guillemin, *Theory of Linear Physical Systems*, John Wiley & Sons, Inc., New York, 1963.
2. A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill Book Company, New York, 1962.
3. S. Mason and H. J. Zimmermann, *Electronic Circuits, Signals and Systems*, John Wiley & Sons, Inc., New York, 1960.
4. J. R. Ragazzini and G. F. Franklin, *Sampled Data Control Systems*, McGraw-Hill Book Company, New York, 1958.
5. H. Freeman, *Discrete-Time Systems*, John Wiley & Sons, Inc., New York, 1965.
6. B. C. Kuo, *Discrete-Data Control Systems*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1970.
7. K. Steiglitz, "The Equivalence of Analog and Digital Signal Processing," *Inform. Control*, Vol. 8, No. 5, Oct. 1965, pp. 455-467.
8. A. V. Oppenheim and D. H. Johnson, "Discrete Representation of Signals," *Proc. IEEE*, Vol. 60, No. 6, June 1972, pp. 681-691.

PROBLEMI

1. Si consideri un sistema lineare arbitrario con ingresso $x(n)$ ed uscita $y(n)$. Mostrare che se $x(n) = 0$ per tutti gli n , allora $y(n)$ deve essere zero per tutti gli n .
2. Per ognuna delle sequenze in fig. P1.2, usare la convoluzione discreta per trovare la risposta all'ingresso $x(n)$ del sistema lineare invariante alla traslazione con risposta all'impulso $h(n)$.

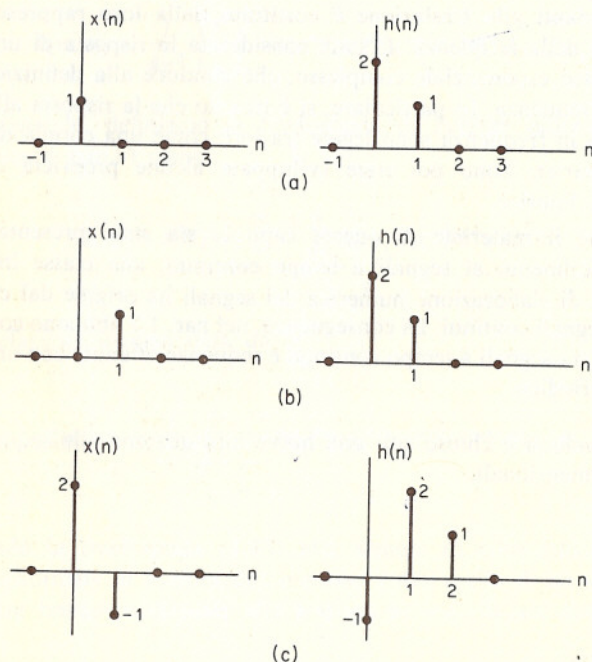


Fig. P1.2

3. Valutare esplicitamente la convoluzione $y(n) = x(n) * h(n)$ delle sequenze

$$h(n) = \begin{cases} \alpha^n, & 0 \leq n < N \\ 0, & \text{altrove} \end{cases}$$

$$x(n) = \begin{cases} \beta^{n-n_0}, & n_0 \leq n \\ 0, & n < n_0 \end{cases}$$

(È possibile ottenere la soluzione in forma chiusa).

4. Sia $e(n)$ una sequenza esponenziale, cioè,

$$e(n) = \alpha^n, \quad \text{per tutti gli } n$$

e siano $x(n)$ e $y(n)$ due sequenze arbitrarie. Mostrare che risulta

$$[e(n)x(n)] * [e(n)y(n)] = e(n)[x(n) * y(n)]$$

5. Siano $x(n)$, $y(n)$ e $w(n)$ tre sequenze arbitrarie. Mostrare che la convoluzione discreta gode delle proprietà
 - (a) commutativa, cioè

$$x(n) * y(n) = y(n) * x(n)$$

- (b) associativa, cioè

$$x(n) * [y(n) * w(n)] = [x(n) * y(n)] * w(n)$$

- (c) distributiva rispetto alla somma, cioè

$$x(n) * [y(n) + w(n)] = x(n) * y(n) + x(n) * w(n)$$

6. Si consideri un sistema lineare a tempo discreto invariante alla traslazione con risposta all'impulso $h(n)$. Se l'ingresso $x(n)$ è una sequenza periodica di periodo N , cioè $x(n) = x(n + N)$, mostrare che anche l'uscita $y(n)$ è una sequenza periodica di periodo N .
7. Se l'uscita di un sistema è la funzione di ingresso moltiplicata per una costante complessa, allora la funzione di ingresso è detta *autofunzione* del sistema.
 - (a) Mostrare che la funzione $x(n) = z^n$, dove z è una costante complessa, è una autofunzione di un sistema lineare invariante alla traslazione.
 - (b) Per mezzo di un controesempio, mostrare che $z^n u(n)$ non è una autofunzione di un sistema lineare invariante alla traslazione.
8. Si sa che la risposta all'impulso di un sistema lineare invariante alla traslazione è nulla eccetto che nell'intervallo $N_0 \leq n \leq N_1$. Si sa che l'ingresso $x(n)$ è zero eccetto che nell'intervallo $N_2 \leq n \leq N_3$. Di conseguenza l'uscita $y(n)$ è anch'essa zero ad esclusione di un certo intervallo $N_4 \leq n \leq N_5$. Determinare N_4 e N_5 in termini di N_0 , N_1 , N_2 e N_3 .
9. Tramite valutazione diretta della somma di convoluzione, determinare la risposta al gradino di un sistema lineare invariante alla traslazione la cui risposta all'impulso è data da

$$h(n) = a^{-n} u(-n), \quad 0 < a < 1$$

10. Si consideri un sistema con una risposta all'impulso $h(n)$ di durata finita tale che

$$h(n) = 0, \quad n < 0, N \leq n, \quad \text{dove } N > 0$$

Mostrare che se $|x(n)| \leq B$ allora un limite sull'uscita è dato da

$$|y(n)| \leq B \sum_{k=0}^{N-1} |h(k)|$$

Mostrare anche che tale limite può essere raggiunto; cioè, determinare una sequenza $x(n)$ con $|x(n)| \leq B$, per la quale, per qualche valore di n ,

$$y(n) = B \sum_{k=0}^{N-1} |h(k)|.$$

11. La causalità di un sistema è stata definita nel par. 1.3. Partendo da questa definizione mostrare che per un sistema lineare invariante alla traslazione, la causalità implica che la risposta all'impulso $h(n)$ è zero per $n < 0$. Mostrare anche che se la risposta all'impulso è zero per $n < 0$, allora il sistema deve necessariamente essere causale.

12. Per ciascuno dei seguenti sistemi, determinare se il sistema è o no (1) stabile, (2) causale, (3) lineare, (4) invariante alla traslazione.

$$\begin{aligned} \text{(a)} \quad T[x(n)] &= g(n)x(n). \\ \text{(b)} \quad T[x(n)] &= \sum_{k=n_0}^n x(k). \\ \text{(c)} \quad T[x(n)] &= \sum_{k=n-n_0}^{n+n_0} x(k). \\ \text{(d)} \quad T[x(n)] &= x(n - n_0). \\ \text{(e)} \quad T[x(n)] &= e^{x(n)}. \\ \text{(f)} \quad T[x(n)] &= ax(n) + b. \end{aligned}$$

Giustificare le risposte.

13. Si consideri un sistema con ingresso $x(n)$ ed uscita $y(n)$. La relazione ingresso-uscita per il sistema è definita dalle seguenti due proprietà: (1) $y(n) - ay(n-1) = x(n)$; (2) $y(0) = 1$.

- (a) Determinare se il sistema è invariante alla traslazione.
 (b) Determinare se il sistema è lineare.
 (c) Si assuma che l'equazione alle differenze (proprietà 1) resti la stessa mentre il valore $y(0)$ è specificato essere nullo. Verificare se le risposte date ai punti (a) e (b) devono essere modificate.

14. Si consideri il sistema lineare a tempo discreto invariante alla traslazione con risposta all'impulso

$$h(n) = \left(\frac{j}{2}\right)^n u(n), \quad \text{dove } j = \sqrt{-1}$$

Determinare la risposta a regime, cioè la risposta, per n grande, alla eccitazione

$$x(n) = [\cos \pi n]u(n)$$

15. In fig. P1.15 è mostrato un sistema a tempo discreto. La trasformazione del sistema $y(n) = T[x(n)]$ è arbitraria e può essere non lineare e tempo variante. La sola proprietà nota del sistema è che esso è ben definito, cioè che l'uscita per un qualsiasi ingresso dato è unica. Si supponga che l'ingresso $x(n)$ sia $x(n) = Ae^{j\omega n}$ e che sia misurato qualche parametro P dell'uscita (ad es. l'ampiezza massima). In generale P sarà una funzione di ω .

Si consideri P per differenti frequenze di eccitazione. Mostrare che P è periodico in ω e determinare il periodo. Un simile risultato sarà ancora vero nel caso a tempo continuo?

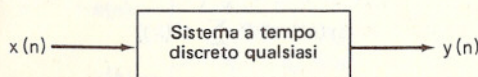


Fig. P1.15

16. Nel par. 1.5 la trasformata di Fourier di una sequenza è stata definita come

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n} \quad (\text{Trasformata di Fourier}) \quad (\text{P1.16-1})$$

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{j\omega})e^{j\omega n} d\omega \quad (\text{Trasformata di Fourier inversa}) \quad (\text{P1.16-2})$$

- (a) Sostituendo la (P1.16-1) nella (P1.16-2) e calcolando l'integrale, verificare che le due relazioni sono l'una l'inversa dell'altra.

- (b) Ripetere la parte (a) sostituendo la (P1.16-2) nella (P1.16-1).

17. Nel par. 1.5 era stato mostrato intuitivamente che

$$Y(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega}) \quad (\text{P1.17-1})$$

dove $Y(e^{j\omega})$, $H(e^{j\omega})$ e $X(e^{j\omega})$ sono le trasformate di Fourier dell'uscita $y(n)$, della risposta all'impulso $h(n)$ e dell'ingresso $x(n)$ di un sistema lineare tempo invariante; cioè,

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)x(k) \quad (\text{P1.17-2})$$

Verificare la (P1.17-1) applicando la trasformata di Fourier alla somma di convoluzione (P1.17-2).

18. (a) Si consideri un sistema lineare tempo invariante con risposta all'impulso $h(n) = \alpha^n u(n)$, dove α è reale e $0 < \alpha < 1$. Se l'ingresso è $x(n) = \beta^n u(n)$, $0 < |\beta| < 1$, determinare l'uscita $y(n)$ nella forma $y(n) = (k_1 \alpha^n + k_2 \beta^n)u(n)$ valutando esplicitamente la somma di convoluzione.
 (b) Valutando esplicitamente le trasformate $X(e^{j\omega})$, $H(e^{j\omega})$ e $Y(e^{j\omega})$ corrispondenti a $x(n)$, $h(n)$ e $y(n)$ specificate nella parte (a), mostrare che

$$Y(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$$

19. Siano $x(n)$ e $X(e^{j\omega})$ una sequenza e la sua trasformata di Fourier. Mostrare che

$$\sum_{n=-\infty}^{\infty} x(n)x^*(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})X^*(e^{j\omega}) d\omega$$

Questa è una forma del teorema di Parseval.

20. Un sistema lineare causale invariante alla traslazione è descritto dall'equazione alle differenze

$$y(n) - ay(n-1) = x(n) - bx(n-1)$$

Determinare il valore di b ($b \neq a$) tale che il sistema sia un sistema passa-tutto, cioè il modulo della sua risposta in frequenza sia una costante indipendente dalla frequenza.

21. Mostrare che la sequenza $[\sin(\pi n/2)]/\pi n$ è quadrato-sommabile ma non assolutamente sommabile.
 22. $f(n)$ e $g(n)$ sono sequenze reali, causali e stabili con trasformate di Fourier $F(e^{j\omega})$ e $G(e^{j\omega})$ rispettivamente. Mostrare che

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{j\omega})G(e^{j\omega}) d\omega = \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{j\omega}) d\omega \right\} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\omega}) d\omega \right\}$$

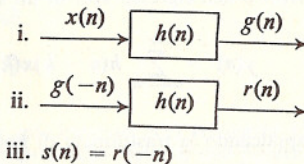
23. Nel progetto di filtri sia analogici che numerici, viene spesso approssimata una caratteristica di ampiezza specificata, senza particolare riguardo alla fase. Per esempio, esistono delle tecniche standard di progetto di filtri passa-basso e passa-banda basate solo sulle caratteristiche di ampiezza.

In molti problemi di filtraggio si desidererebbe idealmente che la caratteristica di fase fosse zero oppure lineare. Per filtri causali è impossibile avere fase zero. Tuttavia, in molte applicazioni di filtraggio numerico non è necessario che la risposta all'impulso del filtro sia nulla per $n < 0$ se l'elaborazione non deve essere fatta in tempo reale.

Una tecnica comunemente usata nel filtraggio numerico quando i dati da filtrare sono di durata finita e immagazzinati, ad esempio, su disco o nastro magnetico, è quella di elaborare i dati facendoli passare prima in un senso e poi in senso opposto nello stesso filtro.

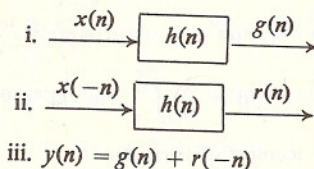
Sia $h(n)$ la risposta all'impulso di un filtro causale con una caratteristica di fase arbitraria. Assumiamo che $h(n)$ sia reale e indichiamo con $H(e^{j\omega})$ la sua trasformata di Fourier. Sia $x(n)$ la sequenza da filtrare. L'operazione di filtraggio è eseguita come segue:

(a) Metodo A:



- (1) Determinare la risposta all'impulso globale $h_s(n)$ che lega $x(n)$ e $s(n)$, e mostrare che ha caratteristica di fase nulla.
- (2) Determinare $|H_s(e^{j\omega})|$ ed esprimerlo in termini di $|H(e^{j\omega})|$ e di $\arg[H(e^{j\omega})]$.

(b) Metodo B: Filtrare $x(n)$ con il filtro $h(n)$ ottenendo $g(n)$. Filtrare anche in senso inverso $x(n)$ con il filtro $h(n)$ ottenendo $r(n)$. L'uscita $y(n)$ è poi presa come la somma di $g(n)$ e $r(-n)$



Questo insieme di più operazioni può essere rappresentato con un filtro, con ingresso $x(n)$, uscita $y(n)$ e risposta all'impulso $h_s(n)$.

- (1) Mostrare che il filtro risultante $h_s(n)$ ha caratteristica di fase nulla.
- (2) Determinare $|H_s(e^{j\omega})|$ ed esprimerlo in termini di $|H(e^{j\omega})|$ e di $\arg[H(e^{j\omega})]$.
- (c) Supponiamo che sia data una sequenza di durata finita, su cui eseguire un'operazione di filtraggio passa-banda con fase nulla. Inoltre, assumiamo che sia dato il filtro passa-banda $h(n)$, con la risposta in frequenza illustrata in fig. P1.23, che ha la caratteristica di ampiezza desiderata ma fase lineare.

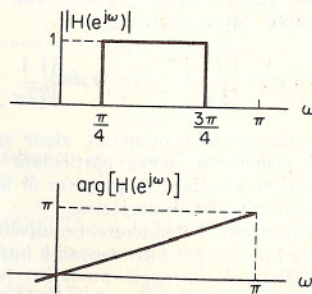


Fig. P1.23

Per ottenere fase nulla possiamo usare sia il metodo (A) che il (B). Determinare e disegnare approssimativamente $|H_1(e^{j\omega})|$ e $|H_2(e^{j\omega})|$. In base a questi risultati, quale metodo usereste per ottenere l'operazione di filtraggio passa-banda desiderata? Spiegare perché. Più in generale, se $h(n)$ ha il modulo desiderato ma caratteristica di fase non lineare, quale metodo è preferibile per ottenere caratteristica di fase nulla?

24. Siano $x(n)$ e $X(e^{j\omega})$ una sequenza e la sua trasformata. Non si assuma che $x(n)$ sia reale o che $x(n)$ sia nulla per $n < 0$. Determinare, in termini di $X(e^{j\omega})$, la trasformata di ognuna delle sequenze seguenti:

- (a) $kx(n)$, con k costante qualsiasi.
- (b) $x(n - n_0)$, con n_0 intero.
- (c) $g(n) = x(2n)$. (Questo caso richiede attenzione!).
- (d) $g(n) = \begin{cases} x(n/2), & \text{per } n \text{ pari} \\ 0, & \text{per } n \text{ dispari.} \end{cases}$
- (e) $x^2(n)$.

25. Nel par. 1.6 abbiamo enunciato alcune proprietà di simmetria della trasformata di Fourier. Tutte queste proprietà discendono in maniera abbastanza diretta dalla definizione di trasformata. Un elenco di alcune delle proprietà enunciate segue più sotto: dimostrare che sono vere. Nel fare la dimostrazione potete usare la definizione della coppia di trasformate data dalle eq. (1.19) e (1.20) nonché qualunque proprietà che precede nella lista. Per esempio, potete usare le proprietà 1 e 2 nel dimostrare la proprietà 3.

Sequenza

1. $x^*(n)$
2. $x^*(-n)$
3. $\text{Re}[x(n)]$
4. $j \text{Im}[x(n)]$
5. $x_e(n)$
6. $x_o(n)$

Trasformata di Fourier

- $X^*(e^{-j\omega})$
- $X^*(e^{j\omega})$
- $X_e(e^{j\omega})$
- $X_o(e^{j\omega})$
- $\text{Re}[X(e^{j\omega})]$
- $j \text{Im}[X(e^{j\omega})]$

26. $x(n)$ e $X(e^{j\omega})$ indicano una sequenza e la sua trasformata di Fourier. Determinare, in termini di $x(n)$, la sequenza corrispondente a

- (a) $X(e^{j(\omega - \omega_0)})$.
- (b) $\text{Re}[X(e^{j\omega})]$.
- (c) $\text{Im}[X(e^{j\omega})]$.

27. Usando le proprietà dimostrate nel probl. 25, mostrare che valgono le seguenti proprietà di simmetria per la trasformata di Fourier $X(e^{j\omega})$ di una sequenza $x(n)$ reale:

$$\begin{aligned} \text{Re}[X(e^{j\omega})] &= \text{Re}[X(e^{-j\omega})] \\ \text{Im}[X(e^{j\omega})] &= -\text{Im}[X(e^{-j\omega})] \\ |X(e^{j\omega})| &= |X(e^{-j\omega})| \\ \arg[X(e^{j\omega})] &= -\arg[X(e^{-j\omega})] \end{aligned}$$

28. Si consideri una sequenza complessa $h(n) = h_r(n) + jh_i(n)$, dove $h_r(n)$ e $h_i(n)$ sono sequenze reali, e sia $H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega})$ la sua trasformata, dove $H_R(e^{j\omega})$ e $H_I(e^{j\omega})$ indicano rispettivamente la parte reale e immaginaria di $H(e^{j\omega})$.

Siano $H_{ER}(e^{j\omega})$ e $H_{OR}(e^{j\omega})$ le parti pari e dispari di $H_R(e^{j\omega})$, e $H_{EI}(e^{j\omega})$ e $H_{OI}(e^{j\omega})$ le parti pari e dispari di $H_I(e^{j\omega})$. Inoltre, indichiamo con $H_A(e^{j\omega})$ e $H_B(e^{j\omega})$ le parti reale e immaginaria della trasformata di $h_r(n)$, e con $H_C(e^{j\omega})$ e $H_D(e^{j\omega})$ le parti reale e immaginaria della trasformata di $h_i(n)$. Esprimere H_A , H_B , H_C e H_D in termini di H_{ER} , H_{OR} , H_{EI} e H_{OI} .

29. Nella fig. P1.29-1 sono illustrate due operazioni che spesso si incontrano nei sistemi per l'elaborazione dei segnali. Il campionatore trasferisce in uscita i cam-

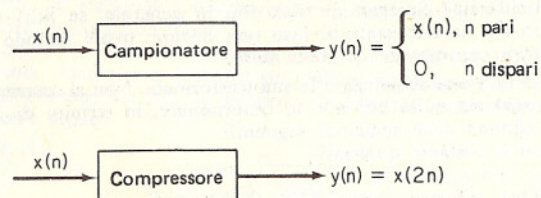


Fig. P1.29-1

pioni con indice pari e pone uguali a zero i campioni della sequenza con indice dispari. Il compressore genera una sequenza che consiste dei soli valori dell'ingresso con indice pari.

In fig. P1.29-2 è mostrato un sistema costituito dalla cascata di un campio-

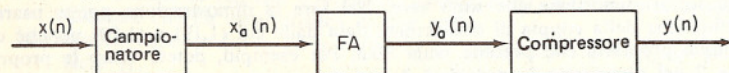


Fig. P1.29-2

natore, un filtro numerico (FA) ed un compressore. Nella fig. P1.29-3 è mostrato un sistema costituito dalla cascata di un compressore e di un filtro numerico (FB).

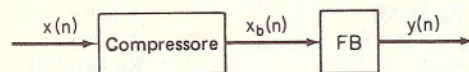


Fig. P1.29-3

I filtri numerici FA e FB sono lineari, causali e invarianti alla traslazione. La risposta in frequenza per FA è $H_A(e^{j\omega}) = 1/[1 - ae^{-j\omega}]$, dove $0 < a < 1$ [cioè $h(n) = a^n u(n)$]. Determinare la risposta in frequenza per FB in modo che i sistemi delle fig. P1.29-2 e P1.29-3 siano equivalenti.

30. Sia $h_a(t)$ la risposta all'impulso di un filtro a tempo continuo lineare tempo-invariante e $h_d(n)$ la risposta al campione unitario di un filtro a tempo discreto lineare invariante alla traslazione.

(a) Se

$$h_a(t) = \begin{cases} e^{-at}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

determinare la risposta in frequenza del filtro analogico e disegnarne il modulo.

- (b) Se $h_d(n) = h_a(nT)$ con $h_a(t)$ come nella parte (a) del problema, determinare la risposta in frequenza del filtro numerico e disegnarne il modulo.
(c) Per un dato valore di a determinare in funzione di T il valore assoluto minimo della risposta in frequenza del filtro numerico.
31. Un'applicazione in cui i filtri numerici sono spesso usati è il filtraggio di dati analogici limitati in banda, come rappresentato in fig. P1.31, dove T rappresenta

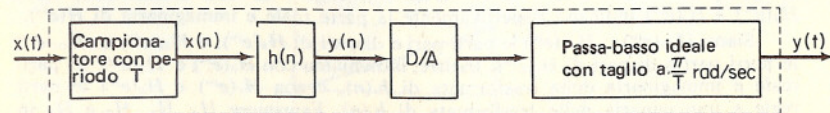


Fig. P1.31

l'intervallo di tempo tra i campioni. (Si assuma T abbastanza piccolo per evitare l'aliasing). Il sistema complessivo che lega $x(t)$ a $y(t)$ è equivalente a un filtro analogico.

- (a) Se $h(n)$ ha una frequenza di taglio di $\pi/8$ rad/sec, e se $1/T = 10$ kHz, qual è la frequenza di taglio del sistema complessivo?
(b) Ripetere la parte (a) per $1/T = 20$ kHz.

32. Come abbiamo accennato nel par. 1.7, una funzione a tempo continuo $x_a(t)$ può in generale essere espressa in termini di un insieme di funzioni di base come

$$x_a(t) = \sum_{k=-\infty}^{\infty} c_k \phi_k(t) \quad (\text{P1.32-1})$$

dove le funzioni $\{\phi_k(t)\}$ sono un insieme di funzioni linearmente indipendenti. I coefficienti c_k possono allora essere pensati come una sequenza che rappresenta il segnale a tempo continuo. Se $x_a(t)$ è a banda limitata, allora, come visto nel par. 1.7, le funzioni $\phi_k(t)$ possono essere scelte come

$$\phi_k(t) = \frac{\sin[(\pi/T)(t - kT)]}{(\pi/T)(t - kT)} \quad (\text{P1.32-2})$$

Un'utile proprietà delle funzioni $\phi_k(t)$ è che siano scelte in modo che la rappresentazione discreta dei segnali a tempo continuo mantenga la convoluzione, cioè, se

$$x_{a1}(t) = \sum_{k=-\infty}^{\infty} c_{1k} \phi_k(t) \quad (\text{P1.32-3a})$$

$$x_{a2}(t) = \sum_{k=-\infty}^{\infty} c_{2k} \phi_k(t) \quad (\text{P1.32-3b})$$

$$x_{a3}(t) = \sum_{k=-\infty}^{\infty} c_{3k} \phi_k(t) \quad (\text{P1.32-3c})$$

con

$$x_{a3}(t) = \int_{-\infty}^{\infty} x_{a1}(\tau) x_{a2}(t - \tau) d\tau \quad (\text{P1.32-3d})$$

allora risulti

$$c_{3k} = \sum_{n=-\infty}^{\infty} c_{1n} c_{2(k-n)} \quad (\text{P1.32-3e})$$

Questo vuol dire quindi che un sistema lineare tempo-invariante a tempo continuo può essere rappresentato da un sistema lineare a tempo discreto invariante alla traslazione.

- (a) Mostrare che se vale la (P1.32-3), allora $\Phi_k(j\Omega)$, cioè la trasformata di Fourier di $\phi_k(t)$, deve soddisfare la relazione

$$\Phi_k(j\Omega) \Phi_n(j\Omega) = \Phi_{k+n}(j\Omega) \quad (\text{P1.32-4})$$

- (b) Mostrare che il risultato della parte (a) implica che $\Phi_k(j\Omega)$ è della forma

$$\Phi_k(j\Omega) = [G(j\Omega)]^{-k}$$

- (c) Mostrare che la trasformata di Fourier delle funzioni (P1.32-2) moltiplicate per $1/T$ soddisfa la (P1.32-4).
(d) Riuscite a trovare altri insiemi di funzioni linearmente indipendenti le cui trasformate soddisfano la (P1.32-4)? Fate qualche esempio.

33. In molti sistemi di comunicazione l'informazione viene trasmessa per mezzo di un insieme di segnali continui distinti, $f_1(t)$, $f_2(t)$, ..., $f_m(t)$. Se tutti i segnali dell'insieme hanno la stessa energia, allora un canale del ricevitore predisposto per ricevere, ad es., $f_1(t)$ può calcolare la quantità $M_r = \int_{-\infty}^{+\infty} s(t)f_1(t) dt$, dove $s(t)$, il segnale ricevuto, è uno dei segnali dell'insieme.

(a) Mostrare, usando la disuguaglianza di Schwarz, che M_r è massimo quando $s(t) = f_1(t)$. (Notare che qui stiamo parlando di funzioni a tempo continuo). Questo significa che per determinare quale segnale è stato inviato occorre calcolare M_r per ogni r . Il valore massimo ottenuto corrisponde al segnale che è stato trasmesso.

(b) Si supponga che, in un particolare sistema, tutti i segnali dell'insieme siano a banda limitata, e che la frequenza massima sia f_0 Hz. Si desidera realizzare il ricevitore in forma numerica, vale a dire che i segnali ricevuti $s(t)$ e $f_1(t)$ sono campionati con periodo di campionamento T per ottenere le sequenze $s(n)$ e $f_1(n)$ e l'uscita del canale viene calcolata come

$$M_r = \sum_{n=-\infty}^{\infty} s(n)f_1(n)$$

Qual è il minimo periodo di campionamento T necessario in relazione a f_0 per rendere questo sistema equivalente a quello continuo? Le ipotesi più ovvie sono che $1/T$ debba essere $f_0/4$ o $f_0/2$ o f_0 , $2f_0$, o $4f_0$. Giustificare completamente la risposta.

(c) Si supponga di aver calcolato M come specificato nella parte (b) con T lungo il doppio del necessario. Esistono condizioni per cui M_r può ancora avere un massimo quando $s(n) = f_1(n)$? Meditare con cura prima di rispondere.

34. Nel par. 1.7 abbiamo derivato la formula d'interpolazione (1.32), che esprime un segnale a tempo continuo limitato in banda in termini dei suoi campioni. In particolare, se $x_a(t)$ indica un segnale a tempo continuo limitato in banda, è stato dimostrato che

$$x_a(t) = \sum_{k=-\infty}^{\infty} x_a(kT) \frac{\sin[(\pi/T)(t - kT)]}{(\pi/T)(t - kT)} \quad (\text{P1.34-1})$$

dove T è minore di 2π diviso per la larghezza di banda di $x_a(t)$.

Spesso, per convertire una sequenza in un segnale a tempo continuo, la sequenza è dapprima convertita in una funzione continua a tratti e poi filtrata con un passa-basso come mostrato in fig. P1.34. Determinare la risposta in frequenza del filtro passa-basso necessario per riottenere $x_a(t)$.

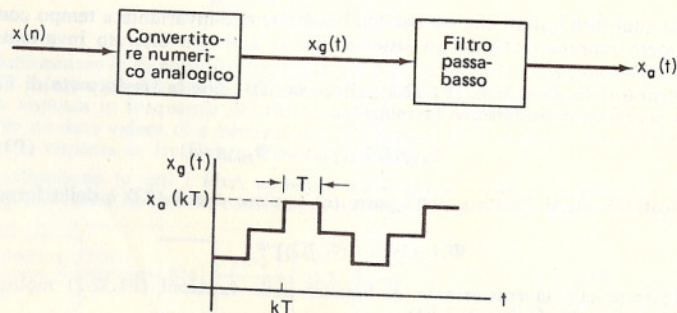


Fig. P1.34

35. Si consideri un segnale analogico

$$x_a(t) = s_a(t) + \alpha s_a(t - T)$$

Assumiamo che la trasformata di Fourier di $x_a(t)$ sia a banda limitata in modo che $X_a(j\Omega) = 0$ per $|\Omega| > \pi/T$ e che $x_a(t)$ sia campionato alla frequenza di Nyquist per ottenere la sequenza

$$x(n) = x_a(nT)$$

Trovare la risposta all'impulso di un sistema a tempo discreto tale che

$$x(n) = \sum_{k=-\infty}^{\infty} s(k)h(n - k)$$

dove $s(n) = s_a(nT)$.

36. $u(m, n)$ e $\delta(m, n)$ indicano rispettivamente le sequenze bidimensionali gradino unitario e campione unitario.

(a) Esprimere $u(m, n)$ in termini di $\delta(m, n)$.

(b) Esprimere $\delta(m, n)$ in termini di $u(m, n)$.

37. Si consideri un sistema bidimensionale lineare invariante alla traslazione con ingresso $x(m, n)$, uscita $y(m, n)$ e risposta all'impulso $h(m, n)$. Mostrare che se $x(m, n)$ e $h(m, n)$ sono entrambe separabili, allora anche $y(m, n)$ è separabile.

38. Sia $x(m, n)$ una sequenza bidimensionale e $X(e^{j\omega_1}, e^{j\omega_2})$ la sua trasformata di Fourier. Mostrare che, se $x(m, n)$ è separabile, anche $X(e^{j\omega_1}, e^{j\omega_2})$ è separabile.

39. Usando un ragionamento analogo a quello seguito nel par. 1.3 per il caso monodimensionale, mostrare che condizione necessaria e sufficiente per la stabilità di un sistema bidimensionale lineare invariante alla traslazione con risposta all'impulso $h(k, r)$ è che sia

$$\sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} |h(k, r)| < \infty$$

40. Nel probl. 19 è stato derivato il teorema di Parseval per sequenze monodimensionali. Derivare la relazione corrispondente per sequenze bidimensionali.

41. Determinare la risposta in frequenza del filtro bidimensionale avente risposta all'impulso

$$h(m, n) = \begin{cases} 1, & |m| < M \text{ e } |n| < N \\ 0, & \text{altrove} \end{cases}$$

42. Determinare la risposta all'impulso del filtro passa-basso la cui risposta in frequenza per $|\omega_1|$ e $|\omega_2|$ minori di π è data da

$$H(e^{j\omega_1}, e^{j\omega_2}) = \begin{cases} 1, & |\omega_1| < a \text{ e } |\omega_2| < b \\ 0, & \text{altrove} \end{cases}$$

43. Determinare la risposta all'impulso del filtro passa-basso a simmetria circolare la cui risposta in frequenza per $|\omega_1|$ e $|\omega_2|$ minori di π è data da

$$H(e^{j\omega_1}, e^{j\omega_2}) = \begin{cases} 1, & \sqrt{\omega_1^2 + \omega_2^2} < A \\ 0, & \text{altrove} \end{cases}$$

2. LA TRASFORMATTA z

2.0 INTRODUZIONE

Nella teoria dei sistemi a tempo continuo la trasformata di Laplace può essere considerata come una generalizzazione della trasformata di Fourier. In modo simile è possibile generalizzare la trasformata di Fourier per segnali e sistemi a tempo discreto, dando luogo a quella che viene comunemente chiamata trasformata z . La trasformata z gioca un ruolo importante nell'analisi e nella rappresentazione dei sistemi lineari a tempo discreto e invarianti alla traslazione. In questo capitolo definiremo la rappresentazione di una sequenza per mezzo della trasformata z e studieremo in dettaglio come le proprietà di una sequenza sono collegate alle proprietà della sua trasformata z .

Nel discutere la trasformata z faremo ricorso a una quantità di risultati propri della teoria delle variabili complesse. Nell'applicare questi risultati ci sforzeremo di essere precisi, senza pretendere tuttavia un elevato grado di rigore matematico.

2.1 LA TRASFORMATTA z

La trasformata z , $X(z)$, di una sequenza $x(n)$ è definita come

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (2.1)$$

dove z è una variabile complessa. Talora risulterà conveniente indicare la trasformata z di una sequenza $x(n)$ come $\mathcal{Z}[x(n)]$. In alcuni contesti è utile chiamare la trasformata z definita dalla (2.1) col nome di *trasformata z bilatera* e considerare anche la *trasformata z unilatera* definita come

$$X_I(z) = \sum_{n=0}^{\infty} x(n)z^{-n}$$

Chiaramente, se $x(n) = 0$ per $n < 0$, le trasformate z unilatera e bilatera sono equivalenti, ma non vale il viceversa. In questo libro non avremo occasione di utilizzare la trasformata unilatera, ma considereremo soltanto la trasformata z bilatera definita dalla (2.1). Una trattazione esauriente della trasformata z unilatera può essere trovata in una quantità di testi (v., per es., [1,2]).

Esprimendo la variabile complessa z in forma polare come $z = re^{j\omega}$, si può dare alla (2.1) una interpretazione in termini della trasformata di Fourier definita nel precedente capitolo. Specificamente, con z espressa in questa forma, la (2.1) diventa

$$X(re^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)(re^{j\omega})^{-n}$$

oppure

$$X(re^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)r^{-n}e^{-j\omega n} \quad (2.2)$$

Pertanto, concordemente con la (2.2), la trasformata z di $x(n)$ può essere interpretata come la trasformata di Fourier di $x(n)$ moltiplicata per una sequenza esponenziale. Per $r = 1$, cioè per $|z| = 1$, la trasformata z è uguale alla trasformata di Fourier della sequenza.

Come abbiamo visto nel cap. 1, la serie di potenze che rappresenta la trasformata di Fourier non converge per tutte le sequenze. Analogamente, la trasformata z non converge per tutte le sequenze o per tutti i valori di z . Per ogni sequenza assegnata l'insieme dei valori di z per cui la trasformata z converge si chiama *regione di convergenza*. Come stabilito nel par. 1.5, la convergenza uniforme della trasformata di Fourier richiede che la sequenza sia assolutamente sommabile. Se ciò si applica alla (2.2) si richiede che sia

$$\sum_{n=-\infty}^{\infty} |x(n)r^{-n}| < \infty \quad (2.3)$$

Dovrebbe essere chiaro dalla (2.3) che, a causa della moltiplicazione della sequenza per l'esponenziale reale r^{-n} , è possibile che la trasformata z converga anche se non converge la trasformata di Fourier. Per esempio, la sequenza $x(n) = u(n)$ non è assolutamente sommabile, e di conseguenza la trasformata di Fourier non converge. Tuttavia, $r^{-n}u(n)$ è assolutamente sommabile se $|r| > 1$, e di conseguenza la trasformata z del gradino unitario esiste con regione di convergenza $1 < |z| < \infty$.

In generale la serie di potenze della (2.1) convergerà in una regione anulare del piano z ,

$$R_{x-} < |z| < R_{x+} \quad (2.4)$$

Dove in generale R_{x-} può essere piccolo fino ad annullarsi ed R_{x+} può essere grande fino all'infinito. Per esempio, la regione di convergenza della trasformata z della sequenza $x(n) = u(n)$ è definita da $R_{x-} = 1$, $R_{x+} = \infty$.

Una serie di potenze della forma (2.1) è una *serie di Laurent*. Pertanto, nello studio della trasformata z è possibile usare una quantità di eleganti e potenti teoremi propri della teoria delle funzioni complesse (presentati, per es., in [3]). Una serie di Laurent, e quindi la trasformata z , è una funzione analitica in ogni punto interno alla regione di convergenza, e pertanto la trasformata z e tutte le sue derivate devono essere funzioni continue di z internamente alla regione di convergenza.

Nel par. 1.5 abbiamo visto che esistono alcune sequenze che non sono assolutamente sommabili ma hanno energia finita e che in quei casi si può affermare che la trasformata di Fourier esiste se si accetta la convergenza nel senso che l'errore quadratico medio tende a zero. Un esempio di tali sequenze è la risposta all'impulso di un filtro ideale passa-basso. In questo caso la trasformata z non esiste, in conseguenza del fatto che per il filtro ideale passa-basso la trasformata di Fourier non è una funzione continua e quindi non è analitica. In quei casi in cui esiste la trasformata di Fourier ma non la trasformata z , possiamo ancora considerare la trasformata di Fourier come la trasformata z valutata per $|z| = 1$, sebbene questo, a rigore, non sia corretto.

Una classe importante di trasformate z è quella per cui $X(z)$ è una funzione razionale, cioè un rapporto di polinomi in z . Le radici del polinomio numeratore sono quei valori di z per cui $X(z) = 0$ e sono chiamati *zeri di $X(z)$* . Quei valori di z per cui $X(z)$ è infinita sono chiamati *poli di $X(z)$* . I poli di $X(z)$ per valori finiti di z sono le radici del polinomio denominatore. Inoltre possono aversi poli anche in $z = 0$ o in $z = \infty$. Per le trasformate z razionali esistono diverse importanti relazioni fra la posizione dei poli di $X(z)$ e la regione di convergenza della trasformata z . Chiaramente non possono esserci poli di $X(z)$ internamente alla regione di convergenza, in quanto la trasformata z non converge in corrispondenza di un polo. Inoltre, come vedremo più avanti, la regione di convergenza è limitata da poli.

Spesso è conveniente rappresentare la trasformata z graficamente per mezzo della posizione dei poli e degli zeri nel piano z . Si consideri per esempio la sequenza

$$x(n) = a^n u(n)$$

La trasformata z è data da

$$\begin{aligned} X(z) &= \sum_{n=-\infty}^{\infty} a^n u(n) z^{-n} \\ &= \sum_{n=0}^{\infty} (az^{-1})^n \end{aligned}$$

che converge a

$$X(z) = \frac{1}{1 - az^{-1}}, \quad \text{per } |z| > |a|$$

Riscrivendo $X(z)$ come un rapporto di polinomi in z , vediamo che $X(z)$ ha uno zero in $z = 0$ e un polo in $z = a$. Ciò è rappresentato nel piano z nella fig. 2.1, dove lo zero è indicato con o e il polo con x . La regione di

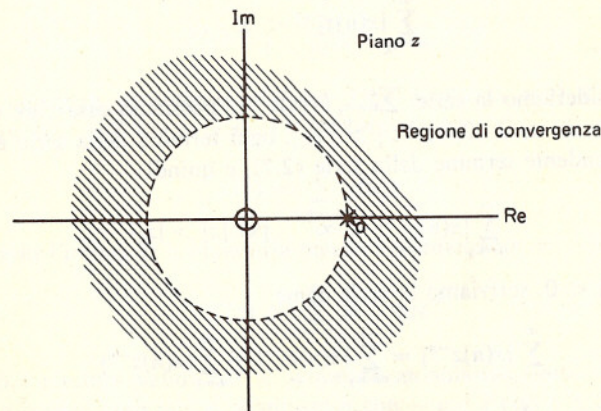


Fig. 2.1 Diagramma di poli e zeri e regione di convergenza nel piano z per la trasformata z della sequenza $a^n u(n)$.

convergenza è indicata dalla regione tratteggiata e include l'intero piano z per $|z| > a$.

Le proprietà della sequenza $x(n)$ determinano la regione di convergenza di $X(z)$. Per vedere come ciò avviene è utile considerare alcuni casi speciali.

1. *Sequenze a lunghezza finita.* Supponiamo che siano diversi da zero solo un numero finito di valori della sequenza, così che

$$X(z) = \sum_{n=n_1}^{n_2} x(n)z^{-n} \quad (2.5)$$

dove n_1 ed n_2 sono numeri interi finiti. La convergenza di questa espressione richiede semplicemente che sia $|x(n)| < \infty$ per $n_1 \leq n \leq n_2$. Inoltre z può assumere tutti i valori ad eccezione di $z = \infty$ se $n_1 < 0$ e di $z = 0$ se $n_2 > 0$. Pertanto le sequenze a lunghezza finita hanno una regione di convergenza che è almeno $0 < |z| < \infty$, e può includere sia $z = 0$ che $z = \infty$.

2. *Sequenze monolateri destre.* Una sequenza $x(n)$ è monolatera destra se $x(n) = 0$ per $n < n_1$. La trasformata z di una tale sequenza è

$$X(z) = \sum_{n=n_1}^{\infty} x(n)z^{-n} \quad (2.6)$$

La regione di convergenza della serie scritta sopra è l'esterno di un cerchio. Per verificarlo, supponiamo che la serie sia assolutamente convergente per $z = z_1$, così che

$$\sum_{n=n_1}^{\infty} |x(n)z_1^{-n}| < \infty \quad (2.7)$$

Se ora consideriamo la serie $\sum_{n=n_1}^{\infty} |x(n)z^{-n}|$, possiamo osservare che, nell'ipotesi che sia $n_1 \geq 0$, se $|z| > |z_1|$, ogni termine della serie è minore del corrispondente termine della serie (2.7), e quindi

$$\sum_{n=n_1}^{\infty} |x(n)z^{-n}| < \infty \quad \text{per } |z| > |z_1|$$

Se invece $n_1 < 0$, scriviamo la serie come

$$\sum_{n=n_1}^{\infty} |x(n)z^{-n}| = \sum_{n=n_1}^{-1} |x(n)z^{-n}| + \sum_{n=0}^{\infty} |x(n)z^{-n}| \quad (2.8)$$

La prima serie nel secondo membro della (2.8) è finita per ogni valore finito di z . La seconda serie, in base allo stesso ragionamento fatto prima, converge per $|z| > |z_1|$. Quindi, se R_{x-} è il più piccolo valore di $|z|$ per il quale la serie (2.6) converge, si ha che la serie converge per

$$R_{x-} < |z|$$

con l'eccezione di $z = \infty$ se $n_1 < 0$ ¹. Pertanto le sequenze monolateri destre hanno una regione di convergenza che è l'esterno di un cerchio con

¹ Notiamo che R_{x-} è il raggio di convergenza della serie di potenze negative di z nella trasformata z della sequenza $x(n)$. Ciò rende conto dell'uso da noi fatto dell'indice $x-$.

raggio R_{x-} . Osserviamo che se $n_1 \geq 0$, e quindi la sequenza è causale, la trasformata z converge in $z = \infty$. Viceversa, se $n_1 < 0$, essa non convergerà in $z = \infty$. Pertanto, se la regione di convergenza della trasformata z è l'esterno di un cerchio, la sequenza è monolatera destra. Inoltre se tale regione include $z = \infty$, la sequenza è causale.

Dalla (2.7) notiamo anche che, poiché la serie converge, ogni termine è limitato e quindi esiste una costante finita A tale che

$$|x(n)z_1^{-n}| < A, \quad n \geq n_1 \quad (2.9)$$

Scrivendo $|z_1|$ come $|z_1| = r$, un numero positivo maggiore di R_{x-} , si ha

$$|x(n)| < Ar^n, \quad n \geq n_1$$

e quindi per convergere la sequenza non può crescere più velocemente di un esponenziale per $n \rightarrow \infty$. Se la regione di convergenza di $x(n)$ si estende internamente al cerchio unitario in modo che r può essere scelto minore di uno, allora, per $n \rightarrow \infty$, $x(n)$ deve tendere a zero velocemente almeno come un esponenziale.

ESEMPIO. Un esempio di sequenza monolatera destra è la sequenza $x(n) = a^n u(n)$, che, come abbiamo visto in precedenza, ha per trasformata z

$$X(z) = \frac{1}{1 - az^{-1}}, \quad |z| > |a| \quad (2.10)$$

3. *Sequenze monolateri sinistre.* Una sequenza $x(n)$ è monolatera sinistra se $x(n) = 0$ per $n > n_2$. La trasformata z è

$$X(z) = \sum_{n=-\infty}^{n_2} x(n)z^{-n} \quad (2.11)$$

Cambiando l'indice di sommatoria con la sostituzione $n = -m$, otteniamo

$$X(z) = \sum_{m=-n_2}^{\infty} x(-m)z^m$$

Pertanto i risultati validi per le sequenze monolateri destre si applicano anche in questo caso pur di sostituire n con $-n$ e z con z^{-1} . Si può dimostrare che la regione di convergenza è l'interno di un cerchio, $|z| > R_{x+}$, con l'eccezione di $z = 0$ se $n_2 > 0$ ². Se la trasformata z di una sequenza monolatera sinistra converge in $z = 0$, allora la sequenza è nulla per $n \geq 0$. Ne segue anche che se $X(z)$ converge per $|z| = r$, allora

$$|x(n)| < Ar^n, \quad n \leq n_2$$

dove A è una costante finita. Per convergere, quindi, la sequenza non può crescere più velocemente di un esponenziale per $n \rightarrow -\infty$. Se la regione

² In questo caso usiamo l'indice $x+$ per indicare il raggio di convergenza della serie di potenze positive di z nella trasformata z della sequenza $x(n)$.

di convergenza include il cerchio unitario, $x(n)$ deve tendere a zero per $n \rightarrow -\infty$.

ESEMPIO. Come esempio di una sequenza monolatera sinistra consideriamo $x(n) = -b^n u(-n-1)$. La trasformata z è

$$\begin{aligned} X(z) &= \sum_{n=-\infty}^{-1} -b^n z^{-n} \\ &= \sum_{n=1}^{\infty} -b^{-n} z^n \\ &= 1 - \sum_{n=0}^{\infty} b^{-n} z^n \end{aligned}$$

La serie converge se $|b^{-1}z| < 1$, cioè $|z| < |b|$, nel qual caso

$$X(z) = 1 - \frac{1}{1 - b^{-1}z} = \frac{z}{z - b}, \quad \text{per } |z| < |b| \quad (2.12)$$

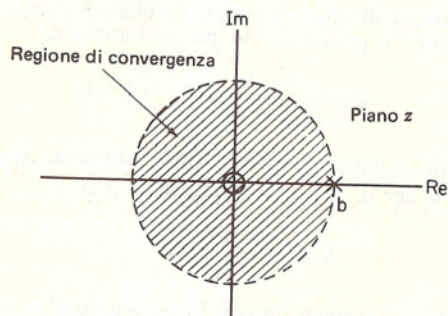


Fig. 2.2 Diagramma di poli e zeri e regione di convergenza nel piano z per la trasformata z della sequenza $-b^n u(-n-1)$.

Il polo e lo zero e la regione di convergenza di $X(z)$ sono indicate in fig. 2.2. Si noti che se $b = a$, la funzione $X(z)$ nella (2.12) è identica a quella della (2.10) dell'esempio precedente. Questo dimostra il fatto estremamente importante che per specificare la trasformata z di una sequenza occorre conoscere non soltanto la funzione $X(z)$, ma anche la regione di convergenza.

4. Sequenze bilatere. Bilatera è una sequenza che si estende da $n = -\infty$ a $n = +\infty$. In generale possiamo scrivere

$$\begin{aligned} X(z) &= \sum_{n=-\infty}^{\infty} x(n) z^{-n} \\ &= \sum_{n=0}^{\infty} x(n) z^{-n} + \sum_{n=-\infty}^{-1} x(n) z^{-n} \end{aligned} \quad (2.13)$$

La prima serie è monolatera destra e converge per $R_{x-} < |z|$; la seconda serie è monolatera sinistra e converge per $|z| < R_{x+}$. Se $R_{x-} < R_{x+}$, esiste una regione comune di convergenza della forma

$$R_{x-} < |z| < R_{x+} \quad (2.14)$$

Se $R_{x-} > R_{x+}$, non esiste una regione comune di convergenza, e quindi la serie (2.13) non converge. Se vale la regione di convergenza espressa in (2.14), la sequenza $x(n)$ non può crescere più velocemente di un esponenziale in entrambe le direzioni, e se

$$R_{x-} < 1 < R_{x+}$$

la sequenza tende a zero esponenzialmente in entrambe le direzioni. Tali sequenze hanno sia una trasformata di Fourier che una trasformata z .

ESEMPIO. Si consideri la sequenza

$$x(n) = \begin{cases} a^n, & n \geq 0 \\ -b^n, & n \leq -1 \end{cases} \quad (2.15)$$

dove $|a| < |b|$. Usando i risultati dei precedenti due esempi,

$$X(z) = \frac{z}{z - a} + \frac{z}{z - b} = \frac{z(2z - a - b)}{(z - a)(z - b)} \quad (2.16)$$

dove la regione di convergenza è

$$|a| < |z| < |b| \quad (2.17)$$

I poli e gli zeri e la regione di convergenza sono mostrati nella fig. 2.3. La regione di convergenza è la parte sovrapposta delle regioni tratteggiate.

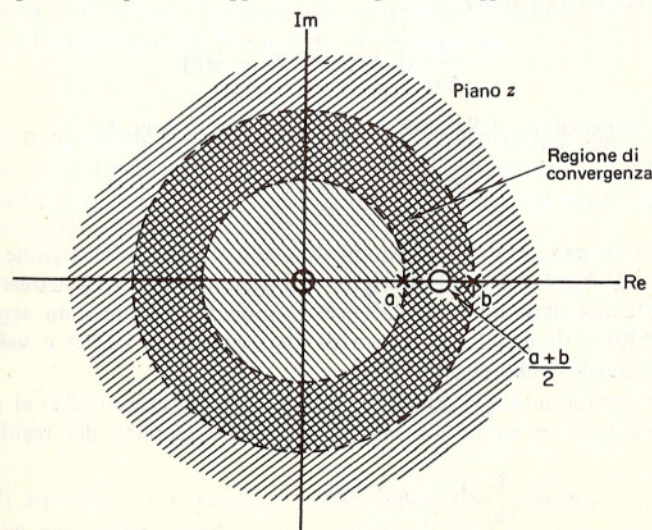


Fig. 2.3 Diagramma di poli e zeri e regione di convergenza per la sequenza $x(n) = a^n u(n) - b^n u(-n-1)$.

2.2 LA TRASFORMATTA Z INVERSA

L'espressione della trasformata z inversa può essere ottenuta utilizzando il *teorema integrale di Cauchy*. Tale teorema afferma che

$$\frac{1}{2\pi j} \oint_C z^{k-1} dz = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases} \quad (2.18)$$

dove C è un contorno percorso in senso antiorario che circonda l'origine.

L'espressione della trasformata z , come è stata definita nel par. 2.1, è data da

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (2.19)$$

Moltiplicando entrambi i membri della (2.19) per z^{k-1} e integrando lungo un percorso che includa l'origine e sia contenuto interamente nella regione di convergenza di $X(z)$, si ottiene

$$\frac{1}{2\pi j} \oint_C X(z)z^{k-1} dz = \frac{1}{2\pi j} \oint_C \sum_{n=-\infty}^{\infty} x(n)z^{-n+k-1} dz \quad (2.20)$$

Integrando termine a termine la sommatoria al secondo membro della (2.20) (il che è giustificato se la serie è convergente) si ottiene

$$\frac{1}{2\pi j} \oint_C X(z)z^{k-1} dz = \sum_{n=-\infty}^{\infty} x(n) \frac{1}{2\pi j} \oint_C z^{-n+k-1} dz \quad (2.21)$$

che per la (2.18) diventa

$$\frac{1}{2\pi j} \oint_C X(z)z^{k-1} dz = x(k)$$

Perciò, l'espressione della trasformata z inversa è data da

$$x(n) = \frac{1}{2\pi j} \oint_C X(z)z^{n-1} dz \quad (2.22)$$

dove C è un percorso antiorario chiuso che è situato nella regione di convergenza di $X(z)$ e circonda l'origine del piano z . Occorre mettere in evidenza che nel derivare la (2.22) non si sono fatte ipotesi sui segni di k nella (2.20) e di n nella (2.22), e di conseguenza la (2.22) è valida per valori sia positivi che negativi di n .

Per trasformate z razionali gli integrali della forma (2.22) si possono spesso valutare in modo conveniente usando il teorema dei residui, cioè

$$\begin{aligned} x(n) &= \frac{1}{2\pi j} \oint_C X(z)z^{n-1} dz \\ &= \sum [\text{residui di } X(z)z^{n-1} \text{ nei poli interni a } C] \end{aligned} \quad (2.23)$$

In generale, se $X(z)z^{n-1}$ è una funzione razionale di z , può essere espressa come

$$X(z)z^{n-1} = \frac{\psi(z)}{(z - z_0)^s} \quad (2.24)$$

dove $X(z)z^{n-1}$ ha s poli in $z = z_0$ e $\psi(z)$ non ha poli in $z = z_0$. Il residuo di $X(z)z^{n-1}$ in $z = z_0$ è dato da

$$\text{Res} [X(z)z^{n-1} \text{ in } z = z_0] = \frac{1}{(s-1)!} \left[\frac{d^{s-1} \psi(z)}{dz^{s-1}} \right]_{z=z_0} \quad (2.25)$$

In particolare, se si ha solo un polo del primo ordine in $z = z_0$, cioè se $s = 1$, allora

$$\text{Res} [X(z)z^{n-1} \text{ in } z = z_0] = \psi(z_0) \quad (2.26)$$

Come esempio dell'uso della trasformazione inversa, consideriamo la trasformata inversa di

$$X(z) = \frac{1}{1 - az^{-1}} \quad |z| > |a|$$

ricavata in un precedente esempio. Usando la (2.23) otteniamo

$$x(n) = \frac{1}{2\pi j} \oint_C \frac{z^{n-1}}{1 - az^{-1}} dz = \frac{1}{2\pi j} \oint_C \frac{z^n dz}{z - a}$$

dove il percorso di integrazione, C , è una circonferenza di raggio maggiore di a . Per $n \geq 0$, allora, il percorso di integrazione racchiude solo un polo in $z = a$. Di conseguenza, per $n \geq 0$, $x(n)$ è dato da

$$x(n) = a^n, \quad n \geq 0$$

Per $n < 0$, si ha un polo multiplo in $z = 0$, il cui ordine dipende da n . Per $n = -1$, il polo è di primo ordine con un residuo di $-a^{-1}$. Il residuo nel polo in $z = a$ è a^{-1} . Di conseguenza, la somma dei residui è zero e perciò $x(-1) = 0$. Per $n = -2$

$$\text{Res} \left[\frac{1}{z^2(z-a)} \text{ in } z = a \right] = a^{-2}$$

e

$$\text{Res} \left[\frac{1}{z^2(z-a)} \text{ in } z = 0 \right] = -a^{-2}$$

e quindi $x(-2) = 0$. Continuando con questo procedimento si può verificare che per questo esempio $x(n) = 0$ per $n < 0$. Quanto più negativo diventa n , tanto più pesante diventa il calcolo del residuo nel polo multi-

plo in $z = 0$. Sebbene la (2.22) sia valida per tutti gli n , il suo uso per $n < 0$ è spesso fonte di complicazioni a causa dei poli multipli in $z = 0$.

Ciò può essere evitato modificando la (2.22) per mezzo di un cambiamento di variabili, rendendola di facile applicazione per $n < 0$. Specificamente, si consideri il cambiamento di variabili $z = p^{-1}$, cosicché la (2.22) diventa

$$x(n) = \frac{-1}{2\pi j} \oint_{C'} X\left(\frac{1}{p}\right) p^{-n+1} p^{-2} dp \quad (2.27)$$

Si osservi che dal momento che il percorso nella (2.22) è antiorario, il percorso nell'espressione precedente è orario. Moltiplicando per -1 per invertire la direzione del percorso, il cambiamento di variabili introdotto porta allora all'espressione

$$x(n) = \frac{1}{2\pi j} \oint_{C'} X\left(\frac{1}{p}\right) p^{-n-1} dp \quad (2.28)$$

Se il percorso C nella (2.22) è una circonferenza di raggio r nel piano z , allora il percorso C' nella (2.28) è una circonferenza di raggio $1/r$ nel piano p . I poli di $X(z)$ che erano all'esterno del percorso C corrispondono ora a poli di $X(1/p)$ che sono all'interno del percorso C' e viceversa. Si può o meno avere la comparsa di poli e/o zeri addizionali nell'origine e/o all'infinito, ma ciò non è essenziale ai fini del nostro ragionamento. Per l'esempio specifico che abbiamo considerato in precedenza, $x(n)$ diventa ora, a seguito di questo cambiamento di variabili,

$$x(n) = \frac{1}{2\pi j} \oint_{C'} \frac{p^{-n-1}}{1 - ap} dp$$

Il percorso di integrazione C' è ora una circonferenza di raggio minore di $1/a$. Per $n < 0$ non vi sono singolarità interne al percorso di integrazione e pertanto segue facilmente che, per $n < 0$, $x(n) = 0$. Come la (2.22) era poco conveniente (sebbene certamente valida) per valutare $x(n)$ per $n < 0$, questa espressione è altrettanto sconsigliabile (ma ancora valida) per valutare $x(n)$ per $n \geq 0$, a causa dei poli multipli che compaiono nell'origine.

In molti casi la valutazione della (2.22) o (2.28) è inutilmente difficile e complicata. Vedremo nel resto di questo paragrafo alcune particolari tecniche che sono spesso di più facile applicazione.

Serie di potenze. Se si dispone della trasformata z in forma di serie di potenze, possiamo osservare che il valore $x(n)$ della sequenza è semplicemente il coefficiente del termine contenente z^{-n} nella serie di potenze

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

Se $x(z)$ è data come un'espressione in forma chiusa, si può spesso ricavare la corrispondente serie di potenze o rifarsi a un'espansione in serie di potenze precedentemente ricavata.

ESEMPIO. Si consideri la trasformata z

$$X(z) = \log(1 + az^{-1}), \quad |a| < |z| \quad (2.29)$$

Usando l'espansione in serie di potenze per $\log(1 + x)$, otteniamo

$$X(z) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} a^n z^{-n}}{n}$$

Perciò, $x(n)$ risulta essere

$$x(n) = \begin{cases} (-1)^{n+1} \frac{a^n}{n}, & n \geq 1 \\ 0, & n \leq 0 \end{cases} \quad (2.30)$$

Per trasformate z razionali si può ottenere un'espansione in serie di potenze usando la divisione lunga.

ESEMPIO. Si consideri la trasformata z

$$X(z) = \frac{1}{1 - az^{-1}}, \quad |z| > |a| \quad (2.31)$$

Poiché la regione di convergenza è l'esterno di un cerchio, la sequenza è monolaterale destra. Inoltre, poiché $X(z)$ tende ad una costante finita al tendere di z all'infinito, la sequenza è causale. Perciò eseguiamo la divisione in modo da ottenere una serie di potenze in z^{-1} . Effettuando la divisione lunga otteniamo

$$\begin{aligned} & \frac{1 + az^{-1} + a^2z^{-2} + \dots}{1 - az^{-1}} = 1 + az^{-1} + a^2z^{-2} + \dots \\ & \frac{1 - az^{-1}}{1 - az^{-1}} = 1 \\ & \frac{az^{-1}}{1 - az^{-1}} = az^{-1} + a^2z^{-2} + \dots \\ & \frac{a^2z^{-2}}{1 - az^{-1}} = a^2z^{-2} + a^3z^{-3} + \dots \end{aligned}$$

e pertanto si ha

$$x(n) = a^n u(n)$$

ESEMPIO. Come altro esempio possiamo considerare lo stesso rapporto di polinomi della (2.31), ma con una diversa regione di convergenza, cioè

$$X(z) = \frac{1}{1 - az^{-1}}, \quad |z| < |a| \quad (2.32)$$

A causa della regione di convergenza, la sequenza è monolatera sinistra e poiché $X(z)$ in $z = 0$ è finita, la sequenza è zero per $n > 0$. Perciò dividiamo in modo da ottenere una serie di potenze in z come segue:

$$-a + z \left| \frac{-a^{-1}z - a^{-2}z^2 - \dots}{z - a^{-1}z^2} \right| \dots$$

Pertanto, $x(n) = -a^n u(-n-1)$.

Espansione in fratti semplici. Un'altra tecnica che spesso è utile per trasformate z razionali consiste nell'effettuare un'espansione in fratti semplici e identificare la trasformata z inversa di termini più facilmente trattabili. Se $F(x)$ è un rapporto di polinomi nella variabile x con il grado del numeratore minore del grado del denominatore e con poli solo del primo ordine, esso può essere espresso in fratti semplici della forma

$$F(x) = \frac{P(x)}{Q(x)} = \sum_{k=1}^N \frac{A_k}{x - x_k} \quad (2.33)$$

dove gli x_k sono i poli di $F(x)$ e gli A_k sono i residui nei poli; cioè

$$A_k = (x - x_k)F(x)|_{x=x_k} \quad (2.34)$$

Se il grado del numeratore è maggiore del grado del denominatore, allora si aggiunge al secondo membro della (2.33) un polinomio di grado pari al grado del numeratore meno il grado del denominatore. Perciò, se il grado di $P(x)$ è M e il grado di $Q(x)$ è N con $M \geq N$, allora la (2.33) è sostituita da

$$F(x) = B_{M-N}x^{M-N} + B_{M-N-1}x^{M-N-1} + \dots + B_1x + B_0 + \sum_{k=1}^N \frac{A_k}{x - x_k} \quad (2.35)$$

I B_i si possono ottenere semplicemente con la divisione lunga e gli A_k si ottengono ancora per mezzo della (2.34). Se $F(x)$ ha poli multipli, la (2.35) va ulteriormente modificata. In particolare, se $F(x)$ ha un polo di ordine s in $x = x_i$, la (2.35) diventa

$$F(x) = B_{M-N}x^{M-N} + B_{M-N-1}x^{M-N-1} + \dots + B_1x + B_0 + \sum_{k=1}^N \frac{A_k}{x - x_k} + \sum_{i=1}^s \frac{C_i}{(x - x_i)^i}$$

I coefficienti A_k e B_i si ottengono come prima. I coefficienti c_k si ricavano dalla relazione

$$C_k = \frac{1}{(s-k)!} \left\{ \frac{d^{s-k}}{dx^{s-k}} [x - x_i]^s F(x) \right\}_{x=x_i} \quad k=1,2,\dots,s$$

Per applicare l'espansione in fratti semplici alla trasformata z possiamo considerare la trasformata z come un rapporto di polinomi in z o in z^{-1} .

ESEMPIO. Si consideri una sequenza monolatera destra con trasformata z

$$X(z) = \frac{1}{(1 - az^{-1})(1 - bz^{-1})} = \frac{z^2}{(z - a)(z - b)} = \frac{a^{-1}b^{-1}}{(z^{-1} - (1/a))(z^{-1} - (1/b))}$$

Effettuando un'espansione in fratti semplici con $X(z)$ considerata come un rapporto di polinomi in z^{-1} , otteniamo

$$X(z) = \frac{a^{-1}b^{-1}}{(z^{-1} - (1/a))(z^{-1} - (1/b))} = \frac{1}{(b-a)} \frac{1}{(z^{-1} - a^{-1})} + \frac{1}{(a-b)} \frac{1}{(z^{-1} - b^{-1})} \\ = \left(\frac{a}{a-b} \right) \left(\frac{1}{1 - az^{-1}} \right) + \left(\frac{b}{b-a} \right) \left(\frac{1}{1 - bz^{-1}} \right) \quad (2.36)$$

Poiché abbiamo assunto che la sequenza sia monolatera destra, ciascuno dei termini della (2.36) corrisponde ad una sequenza monolatera destra. Tali termini sono trasformate z del primo ordine che possiamo riconoscere dagli esempi precedenti, e perciò si può ricavare immediatamente dalla (2.36)

$$x(n) = \frac{a}{a-b} a^n u(n) + \frac{b}{b-a} b^n u(n)$$

Questo stesso esempio verrà preso in esame nel probl. 4 di questo capitolo, considerando però $X(z)$ come rapporto di polinomi in z anziché in z^{-1} .

È bene osservare che per sequenze monolateri sinistre o bilateri la tecnica di espansione in fratti semplici funziona ugualmente bene, ma occorre fare attenzione alla determinazione di quali poli corrispondono a sequenze monolateri destre e quali a sequenze monolateri sinistre.

2.5 TEOREMI E PROPRIETÀ DELLA TRASFORMATTA z

Nel risolvere problemi di trattamento dei segnali, è importante avere ben chiare e saper usare con facilità le proprietà della trasformata z . In questo paragrafo discuteremo alcune delle più importanti tra queste proprietà. Ulteriori nozioni possono essere trovate in [1] e [2].

2.3.1 Regione di convergenza delle trasformate z razionali

Come è stato indicato precedentemente, per una sequenza con trasformata z razionale la regione di convergenza non può contenere alcun polo ed è limitata da poli o da zero o da infinito. Il fatto che essa non contiene alcun polo segue semplicemente dal fatto che per definizione la trasformata z non converge in corrispondenza di un polo. Per comprendere perché essa è limitata da poli, consideriamo prima il caso di una sequenza monolaterale destra ed assumiamo che i poli siano situati in $a_0, a_1 \dots a_N$, dove a_N ha il modulo più grande. Assumeremo, per semplificare le cose, che tutti i poli siano semplici, nonostante sia possibile generalizzare facilmente tale ipotesi. Allora, per n maggiore di un certo valore n_0 , la sequenza consiste di una somma di esponenziali della forma

$$x(n) = \sum_{k=0}^N A_k (a_k)^n, \quad n > n_0 \quad (2.37)$$

La regione di convergenza è determinata dall'insieme dei valori di z per i quali la sequenza $x(n)z^{-n}$ è assolutamente convergente. Poiché una sequenza monolaterale destra della forma $(a_k)^n z^{-n}$ è assolutamente sommabile per $|z| > |a_k|$ ma non per $|z| \leq |a_k|$, ne segue che la sequenza monolaterale destra (2.37) ha una regione di convergenza definita da $|z| > |a_N|$. Di conseguenza essa è limitata verso l'interno dal polo di modulo massimo e verso l'esterno da infinito. Con identiche considerazioni si vede che per una sequenza monolaterale sinistra la regione di convergenza è limitata all'esterno dal polo con il modulo minimo e verso l'interno da zero. Per sequenze bilatere qualcuno dei poli contribuisce soltanto per $n \geq 0$ ed i restanti soltanto per $n \leq 0$. La regione di convergenza è limitata verso l'interno dal polo con il modulo massimo, il quale contribuisce per $n \geq 0$, ed all'esterno dal polo con il modulo più piccolo, che contribuisce per $n \leq 0$. Per esempio, in fig. 2.4 è mostrata una stessa configurazione di zeri e poli con le quattro possibili scelte per la regione di convergenza. La fig. 2.4 (a) corrisponde ad una sequenza monolaterale destra, la 2.4 (b) ad una sequenza monolaterale sinistra, e le restanti due corrispondono a sequenze bilatere. In generale la regione di convergenza è una regione connessa. Non è possibile per esempio considerare come regione di convergenza l'insieme $|z| < |a|$ e $|z| > |c|$.

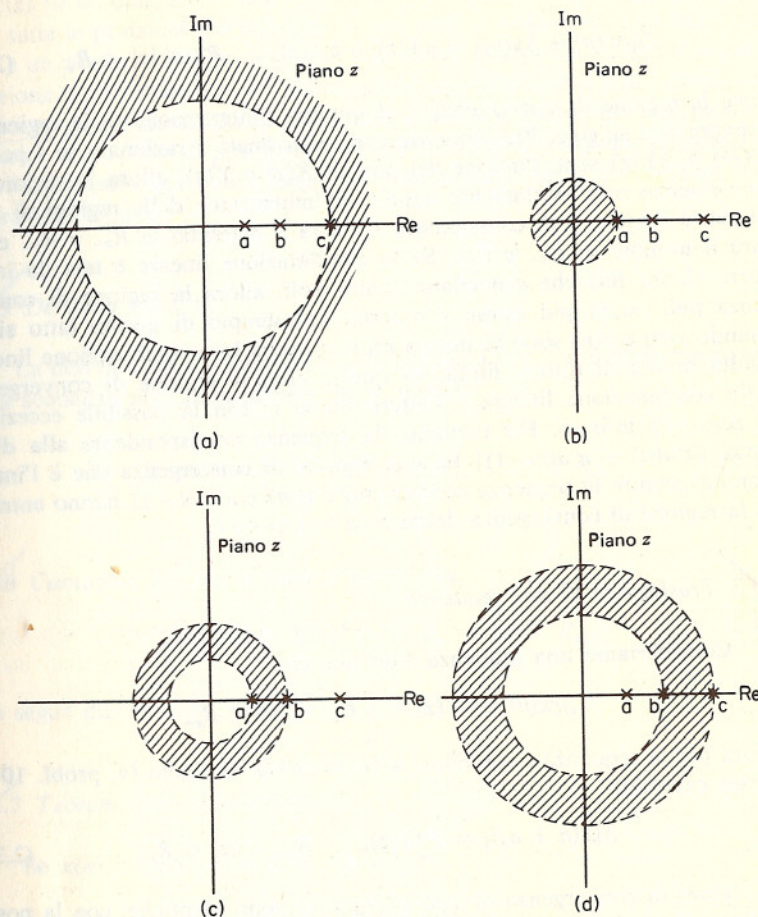


Fig. 2.4 Esempi di quattro trasformate z con la stessa disposizione di poli e zeri, illustranti le differenti possibilità per la regione di convergenza. Ciascuna di esse corrisponde a una sequenza diversa, con (a) corrispondente ad una sequenza monolaterale destra, (b) ad una sequenza monolaterale sinistra, e le rimanenti due a sequenze bilatere.

2.3.2 Linearità

Consideriamo due sequenze $x(n)$ e $y(n)$ con trasformate z , rispettivamente, $X(z)$ e $Y(z)$; cioè

$$\begin{aligned} \mathcal{Z}[x(n)] &= X(z), & R_{x-} < |z| < R_{x+} \\ \mathcal{Z}[y(n)] &= Y(z), & R_{y-} < |z| < R_{y+} \end{aligned}$$

Ne segue che:

$$\mathfrak{Z}[ax(n) + by(n)] = aX(z) + bY(z), \quad R_- < |z| < R_+ \quad (2.38)$$

dove la regione di convergenza è al minimo l'intersezione delle regioni di convergenza singole. Per sequenze con trasformate z razionali, se i poli di $aX(z) + bY(z)$ sono l'unione dei poli di $X(z)$ e $Y(z)$, allora la regione di convergenza sarà esattamente uguale all'intersezione delle regioni di convergenza singole, e di conseguenza R_- sarà il massimo di R_{x-} e R_{y-} e R_+ sarà il minimo di R_{x+} e R_{y+} . Se la combinazione lineare è tale da introdurre alcuni zeri che cancellano alcuni poli, allora la regione di convergenza può essere più estesa. Un semplice esempio di questo fatto si ha quando $x(n)$ e $y(n)$ sono di durata infinita ma la loro combinazione lineare risulta invece di durata finita. In questo caso la regione di convergenza della combinazione lineare è l'intero piano z , con la possibile eccezione di zero e/o infinito. Per esempio, la sequenza corrispondente alla differenza $[a^n u(n) - a^n u(n-1)]$ ha una regione di convergenza che è l'intero piano z , mentre le sequenze componenti $a^n u(n)$ e $a^n u(n-1)$ hanno entrambe la regione di convergenza definita da $|z| > |a|$.

2.3.3 Traslazione di una sequenza

Consideriamo una sequenza $x(n)$ tale che

$$\mathfrak{Z}[x(n)] = X(z), \quad R_{x-} < |z| < R_{x+}$$

Allora per la sequenza i cui valori sono $x(n+n_0)$ abbiamo (v. probl. 10 di questo capitolo)

$$\mathfrak{Z}[x(n+n_0)] = z^{n_0} X(z), \quad R_{x-} < |z| < R_{x+} \quad (2.39)$$

Le regioni di convergenza di $x(n)$ e $x(n+n_0)$ sono identiche, con la possibile eccezione di $z=0$ o $z=\infty$. Per esempio, la sequenza $\delta(n)$ ha una trasformata z che converge dovunque nel piano z , ma la trasformata z di $\delta(n-1)$ non converge per $z=0$ e la trasformata z di $\delta(n+1)$ non converge per $z=\infty$. Come si vede dalla (2.39), per n_0 positivo vengono introdotti degli zeri in $z=0$ e dei poli per $z=\infty$; per n_0 negativo, vengono invece introdotti dei poli nell'origine e degli zeri all'infinito.

2.3.4 Moltiplicazione per una sequenza esponenziale

Se una sequenza $x(n)$ è moltiplicata per una sequenza esponenziale a^n , dove a può essere complesso, allora (v. probl. 10 di questo capitolo)

$$\mathfrak{Z}[a^n x(n)] = X(a^{-1}z), \quad |a| \cdot R_{x-} < |z| < |a| \cdot R_{x+} \quad (2.40)$$

Se $X(z)$ ha un polo in $z=z_1$, allora $X(az^{-1})$ avrà un polo in $z=az_1$. In generale, tutte le posizioni dei poli e degli zeri sono modificate per un fattore a . Se a è un numero reale positivo, ciò può essere interpretato come una compressione o un'espansione del piano z , cioè le posizioni dei poli e degli zeri cambiano lungo linee radiali nel piano z . Se a è complesso con modulo unitario, la modifica di posizione corrisponde ad una rotazione nel piano z , cioè le posizioni dei poli e degli zeri cambiano lungo circonferenze con centro nell'origine.

2.3.5 Derivata di $X(z)$

La derivata della trasformata z , moltiplicata per $-z$, è la trasformata z della sequenza $x(n)$ moltiplicata per n , e cioè

$$\mathfrak{Z}[nx(n)] = -z \frac{dX(z)}{dz}, \quad R_{x-} < |z| < R_{x+} \quad (2.41)$$

2.3.6 Coniugata di una sequenza complessa

$$\mathfrak{Z}[x^*(n)] = X^*(z^*), \quad R_{x-} < |z| < R_{x+} \quad (2.42)$$

Ciò segue direttamente dalla definizione di trasformata z .

2.3.7 Teorema del valore iniziale

Se $x(n)$ è zero per $n < 0$, allora

$$x(0) = \lim_{z \rightarrow \infty} X(z) \quad (2.43)$$

Questo teorema si dimostra facilmente considerando il limite di ogni termine nella serie (2.1) (v. probl. 16 di questo capitolo).

2.3.8 Convoluzione di sequenze

Se $w(n)$ è la convoluzione delle due sequenze $x(n)$ e $y(n)$, allora la trasformata z di $w(n)$ è il prodotto delle trasformate z di $x(n)$ e $y(n)$, cioè, se

$$w(n) = \sum_{k=-\infty}^{\infty} x(k)y(n-k) \quad (2.44)$$

allora $W(z) = X(z)Y(z)$. Per mostrare ciò scriviamo

$$W(z) = \sum_{n=-\infty}^{\infty} \left[\sum_{k=-\infty}^{\infty} x(k)y(n-k) \right] z^{-n}$$

Scambiando l'ordine delle sommatorie si ha

$$W(z) = \sum_{k=-\infty}^{\infty} x(k) \sum_{n=-\infty}^{\infty} y(n-k) z^{-n}$$

e cambiando l'indice della seconda sommatoria da n a $m = n - k$ otteniamo

$$W(z) = \sum_{k=-\infty}^{\infty} x(k) \left[\sum_{m=-\infty}^{\infty} y(m) z^{-m} \right] z^{-k}$$

Di conseguenza, per valori di z all'interno delle regioni di convergenza sia di $Y(z)$ che di $X(z)$, possiamo scrivere

$$W(z) = X(z)Y(z), \quad R_{y-} < |z| < R_{y+}, \quad R_{x-} < |z| < R_{x+} \quad (2.45)$$

dove la regione di convergenza include l'intersezione delle regioni di convergenza di $Y(z)$ e $X(z)$. Se un polo che limita la regione di convergenza di una delle trasformate z è cancellato da uno zero dell'altra, allora la regione di convergenza di $W(z)$ sarà più estesa.

ESEMPIO. Sia $y(n) = a^n u(n)$ e $x(n) = u(n)$. Le corrispondenti trasformate z sono

$$Y(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \frac{1}{1 - az^{-1}}, \quad |z| > |a|$$

e

$$X(z) = \sum_{n=0}^{\infty} z^{-n} = \frac{1}{1 - z^{-1}}, \quad |z| > 1$$

La trasformata della convoluzione è allora

$$\begin{aligned} W(z) &= \frac{1}{(1 - az^{-1})(1 - z^{-1})} \\ &= \frac{z^2}{(z - a)(z - 1)}, \quad |z| > 1 \end{aligned}$$

I poli e gli zeri di $W(z)$ sono mostrati in fig. 2.5 e la regione di convergenza è l'intersezione delle singole regioni di convergenza.

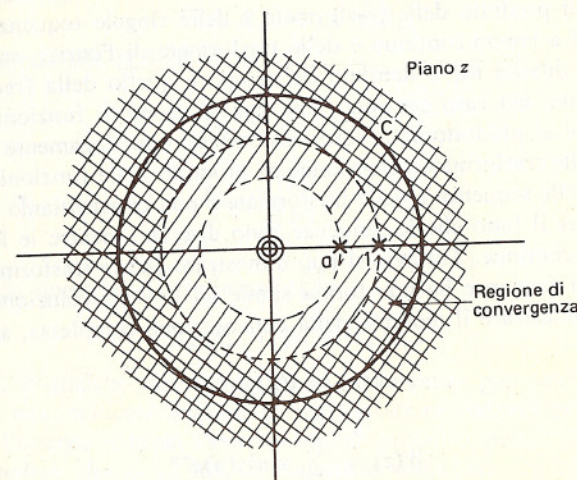


Fig. 2.5 Diagramma di poli e zeri per la trasformata z della convoluzione delle sequenze $u(n)$ e $a^n u(n)$.

La sequenza $w(n)$ può essere ottenuta tramite la formula di inversione

$$w(n) = \frac{1}{2\pi j} \oint_C \frac{z^{n+1} dz}{(z - a)(z - 1)}$$

dove il percorso C è scelto nella regione di convergenza di $w(n)$, come è mostrato in fig. 2.5. Per $n \geq 0$ non ci sono poli in $z = 0$. Perciò, usando il teorema dei residui, si ha

$$\begin{aligned} w(n) &= \frac{(1)^{n+1}}{1 - a} + \frac{a^{n+1}}{a - 1} \\ &= \frac{1 - a^{n+1}}{1 - a}, \quad n \geq 0 \end{aligned}$$

Per $n < -1$, benché ci siano poli in $z = 0$, non è necessario calcolare l'integrale di linea per accorgersi che $w(n)$ deve essere zero per $n < 0$. Notiamo semplicemente che poiché sia $x(n)$ che $y(n)$ sono nulle per $n < 0$, $w(n)$ deve essere anch'essa nulla. Osserviamo anche che, per questo esempio, la regione di convergenza di $W(z)$ era l'intersezione delle regioni di convergenza di $x(n)$ e $y(n)$. Se, invece, avessimo scelto $y(n)$ tale che

$$Y(z) = \frac{1 - z^{-1}}{1 - az^{-1}}, \quad |z| > |a|$$

allora il polo in $z = 1$ sarebbe stato cancellato e la regione di convergenza di $w(n)$ si sarebbe estesa fino al polo in $z = a$.

2.3.9 Teorema della convoluzione complessa

Si è appena visto che la trasformata z della convoluzione di più sequenze è il prodotto delle trasformate z delle singole sequenze. Nel caso dei segnali a tempo continuo e delle trasformate di Fourier, sappiamo che esiste una dualità tra il dominio del tempo e quello della frequenza. Più precisamente, nel caso continuo, una convoluzione di funzioni del tempo corrisponde al prodotto delle loro trasformate, e analogamente una convoluzione delle trasformate corrisponde al prodotto delle funzioni del tempo. Nel caso delle sequenze e delle trasformate z non ci aspettiamo una dualità perfetta, per il fatto che le sequenze sono discrete, mentre le loro trasformate sono continue. Tuttavia si può dimostrare che la trasformata z di un prodotto di sequenze ha una forma simile ad una convoluzione.

Per dimostrare il teorema della convoluzione complessa, sia

$$w(n) = x(n)y(n)$$

e quindi

$$W(z) = \sum_{n=-\infty}^{\infty} x(n)y(n)z^{-n}$$

Ma

$$y(n) = \frac{1}{2\pi j} \oint_{C_1} Y(v)v^{n-1} dv$$

dove C_1 è un percorso chiuso antiorario nella regione di convergenza di $Y(v)$. Allora

$$\begin{aligned} W(z) &= \frac{1}{2\pi j} \sum_{n=-\infty}^{\infty} x(n) \oint_{C_1} Y(v) \left(\frac{z}{v}\right)^{-n} v^{-1} dv \\ &= \frac{1}{2\pi j} \oint_{C_1} \left[\sum_{n=-\infty}^{\infty} x(n) \left(\frac{z}{v}\right)^{-n} \right] v^{-1} Y(v) dv \end{aligned}$$

ovvero

$$W(z) = \frac{1}{2\pi j} \oint_{C_1} X\left(\frac{z}{v}\right) Y(v) v^{-1} dv \quad (2.46a)$$

dove C_1 è un percorso chiuso antiorario nell'intersezione delle regioni di convergenza di $X(z/v)$ e $Y(v)$. $W(z)$ può anche essere espressa come

$$W(z) = \frac{1}{2\pi j} \oint_{C_2} X(v) Y\left(\frac{z}{v}\right) v^{-1} dv \quad (2.46b)$$

dove C_2 è un percorso chiuso nell'intersezione delle regioni di convergenza di $X(v)$ e di $Y(z/v)$.

Per determinare la regione di convergenza associata a $W(z)$, supponiamo

che le regioni di convergenza, rispettivamente, di $X(z)$ e di $Y(z)$ siano

$$X(z): R_{x-} < |z| < R_{x+}$$

$$Y(z): R_{y-} < |z| < R_{y+}$$

Allora riguardo alla (2.46a) si ha

$$R_{y-} < |v| < R_{y+}$$

e

$$R_{x-} < \left| \frac{z}{v} \right| < R_{x+}$$

Combinando queste due espressioni si ha

$$R_{x-}R_{y-} < |z| < R_{x+}R_{y+}$$

Anche qui, in qualche caso la regione di convergenza può essere più estesa di questa, ma includerà sempre la regione sopra definita per estendersi poi verso l'interno o verso l'esterno fino al prossimo polo.

Per vedere che la (2.46b) è realmente simile ad una convoluzione, supponiamo che il percorso di integrazione sia una circonferenza con

$$v = \rho e^{j\theta} \quad \text{e} \quad z = r e^{j\phi}$$

La (2.46b) diventa

$$W(re^{j\phi}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y\left[\frac{r}{\rho} e^{j(\phi-\theta)}\right] X(\rho e^{j\theta}) d\theta \quad (2.47)$$

che ha una forma simile ad una convoluzione. In particolare, a parte i limiti dell'integrale, l'espressione di sopra è identica alla convoluzione di $X(\rho e^{j\theta})$ e $Y(\rho e^{j\theta})$ considerate come funzioni di θ . Notiamo che queste funzioni sono funzioni periodiche di θ e che quindi l'integrazione va fatta soltanto su un periodo. Una convoluzione di questo tipo viene spesso chiamata *convoluzione periodica* e giocherà un ruolo particolarmente importante nel cap. 3.

Nell'applicazione del teorema della convoluzione complessa espresso dalla (2.46a) o (2.46b), una delle difficoltà principali è costituita dal problema di determinare quali poli dell'integrando sono interni al percorso di integrazione e quali sono esterni. Un semplice esempio dell'uso del teorema della convoluzione complessa può servire ad indicare la procedura.

ESEMPIO. Consideriamo $x(n) = a^n u(n)$ e $y(n) = b^n u(n)$. Allora le trasformate z $X(z)$ e $Y(z)$ sono rispettivamente

$$X(z) = \frac{1}{1 - az^{-1}}, \quad |z| > |a|$$

$$Y(z) = \frac{1}{1 - bz^{-1}}, \quad |z| > |b|$$

Sostituendo nella (2.46a), si ottiene

$$W(z) = \frac{1}{2\pi j} \oint_{C_1} \frac{-(z/a)}{(v - z/a)} \frac{1}{v - b} dv$$

L'integrando ha due poli, uno situato in $v = b$ e l'altro in $v = z/a$. Il percorso di integrazione di questa espressione deve essere interno alla regione di convergenza di $Y(v)$, e di conseguenza il polo in $v = b$ dovrà essere interno a tale percorso. Per determinare se anche il polo in $v = z/a$ è interno, teniamo presente che la trasformata z $X(z)$ è valida soltanto per $|z| > |a|$. Pertanto la corrispondente espressione per $X(z/v)$ è valida soltanto per $|z/v| > |a|$. Quindi, se

$$\left| \frac{z}{v} \right| > |a|$$

allora

$$\left| \frac{z}{a} \right| > |v|$$

Di conseguenza il polo deve sempre giacere al di fuori del percorso chiuso di integrazione in v . La posizione dei poli ed il percorso di integrazione sono indicati in fig. 2.6, dove si è supposto che a e b siano reali. Usando il teorema dei residui di Cauchy per calcolare $W(z)$, otteniamo

$$\begin{aligned} W(z) &= \frac{-z/a}{b - z/a} \\ &= \frac{1}{1 - abz^{-1}}, \quad |z| \geq ab \end{aligned}$$

Osserviamo che questa espressione deriva dall'aver tenuto conto solamente del residuo nel polo interno al percorso di integrazione. Si verifica facilmente che se avessimo erroneamente considerato il polo in z/a interno al percorso di integrazione, il risultato del calcolo dell'integrale sarebbe stato indenticamente nullo.

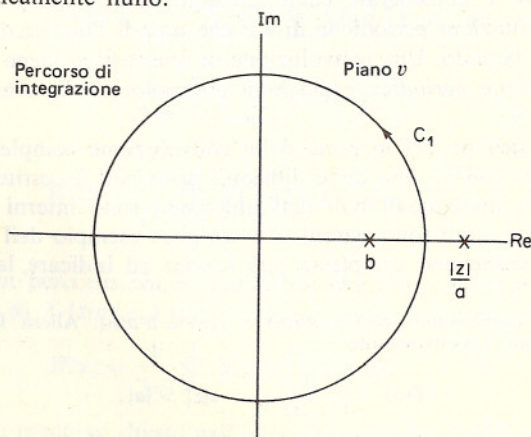


Fig. 2.6 Poli dell'integrando e percorso di integrazione nell'esempio di applicazione del teorema della convoluzione complessa.

2.3.10 Relazione di Parseval

Nel probl. 19 del cap. 1 abbiamo considerato la relazione di Parseval relativa alla trasformata di Fourier. L'estensione di questa relazione alla trasformata z deriva dal teorema della convoluzione complessa. Consideriamo in particolare due sequenze complesse $x(n)$ e $y(n)$. Allora la relazione di Parseval afferma che

$$\sum_{n=-\infty}^{\infty} x(n)y^*(n) = \frac{1}{2\pi j} \oint_C X(v)Y^*(1/v^*)v^{-1} dv \quad (2.48)$$

dove il percorso chiuso di integrazione è preso interno all'intersezione delle regioni di convergenza di $X(v)$ e di $Y^*(1/v^*)$. La relazione di sopra può essere dimostrata definendo una sequenza $w(n)$ come

$$w(n) = x(n)y^*(n) \quad (2.49)$$

e notando che

$$\sum_{n=-\infty}^{\infty} w(n) = W(z)|_{z=1} \quad (2.50)$$

Allora dalla (2.42) e dal teorema della convoluzione complessa segue che

$$W(z) = \frac{1}{2\pi j} \oint_C X(v)Y^*(z^*/v^*)v^{-1} dv$$

Quindi, applicando le (2.49) e (2.50), otteniamo la (2.48). Se $X(z)$ e $Y(z)$ convergono sul circolo unitario, possiamo scegliere $v = e^{j\omega}$, e la (2.48) diventa

$$\sum_{n=-\infty}^{\infty} x(n)y^*(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})Y^*(e^{j\omega}) d\omega \quad (2.51)$$

2.3.11 Riassunto di alcuni teoremi e di alcune proprietà della trasformata z

Nei paragrafi precedenti abbiamo visto e discusso un certo numero di teoremi e di proprietà delle trasformate z . Molte di esse sono utili nelle applicazioni e pertanto sono state riassunte nella tab. 2.1 insieme ad alcune altre proprietà non esplicitamente illustrate nel testo. Le regioni indicate sono incluse nella regione di convergenza, ma in qualche caso questa può essere più estesa.

Tab. 2.1

Sequenza		Trasformata z
1. $x(n)$	$X(z)$	$R_{x-} < z < R_{x+}$
2. $y(n)$	$Y(z)$	$R_{y-} < z < R_{y+}$
3. $ax(n) + by(n)$	$aX(z) + bY(z)$	$\max [R_{x-}, R_{y-}] < z < \min [R_{x+}, R_{y+}]$
4. $x(n + n_0)$	$z^{n_0} X(z)$	$R_{x-} < z < R_{x+}$
5. $a^n x(n)$	$X(a^{-1}z)$	$ a R_{x-} < z < a R_{x+}$
6. $nx(n)$	$-z \frac{dX(z)}{dz}$	$R_{x-} < z < R_{x+}$
7. $x^*(n)$	$X^*(z^*)$	$R_{x-} < z < R_{x+}$
8. $x(-n)$	$X(1/z)$	$1/R_{x+} < z < 1/R_{x-}$
9. $\text{Re} [x(n)]$	$\frac{1}{2}[X(z) + X^*(z^*)]$	$R_{x-} < z < R_{x+}$
10. $\text{Im} [x(n)]$	$\frac{1}{2j}[X(z) - X^*(z^*)]$	$R_{x-} < z < R_{x+}$
11. $x(n) * y(n)$	$X(z)Y(z)$	$\max [R_{x-}, R_{y-}] < z < \min [R_{x+}, R_{y+}]$
12. $x(n)y(n)$	$\frac{1}{2\pi j} \oint_C X(v)Y\left(\frac{z}{v}\right)v^{-1} dv$	$R_{x-}R_{y-} < z < R_{x+}R_{y+}$

2.4. FUNZIONE DI TRASFERIMENTO

Nel cap. 1 abbiamo considerato una descrizione dei sistemi lineari invarianti alla traslazione in termini della trasformata di Fourier della risposta all'impulso. Come si è visto, la trasformata di Fourier della risposta all'impulso corrisponde alla risposta in frequenza del sistema. Inoltre, nel dominio della frequenza la relazione di ingresso-uscita corrisponde semplicemente a una moltiplicazione tra la trasformata di Fourier dell'ingresso e quella della risposta all'impulso.

Più in generale, possiamo descrivere un sistema lineare invariante alla traslazione in termini della trasformata z della risposta all'impulso. Se $x(n)$, $y(n)$ e $h(n)$ indicano rispettivamente l'ingresso, l'uscita e la risposta all'impulso, e $X(z)$, $Y(z)$ e $H(z)$ le loro trasformate z, essendo

$$y(n) = x(n) * h(n)$$

segue dal paragrafo precedente che

$$Y(z) = X(z)H(z) \quad (2.52)$$

Come per la trasformata di Fourier, la relazione di ingresso-uscita per un sistema lineare invariante alla traslazione corrisponde a una moltiplicazione delle trasformate z dell'ingresso e della risposta all'impulso.

La trasformata z della risposta all'impulso viene spesso chiamata funzione di trasferimento. La funzione di trasferimento calcolata sul circolo unitario (cioè per $|z| = 1$) è la risposta in frequenza del sistema.

Nel cap. 1 fu dimostrato che condizione necessaria e sufficiente per la stabilità di un sistema è che la risposta all'impulso $h(n)$ sia assolutamente sommabile. La regione di convergenza della trasformata z è definita come quei valori di z per cui $h(n)z^{-n}$ è assolutamente sommabile. Di conseguenza, se la regione di convergenza della funzione di trasferimento comprende il circolo unitario, il sistema è stabile, e viceversa. Di più, possiamo affermare che per un sistema stabile e causale la regione di convergenza comprenderà il circolo unitario e tutto il piano z esterno al cerchio unitario, incluso $z = \infty$.

Quando il sistema è descrivibile con un'equazione alle differenze lineare a coefficienti costanti, la funzione di trasferimento è un rapporto di polinomi. Per vederlo, consideriamo un sistema per cui l'ingresso e l'uscita soddisfano l'equazione alle differenze di ordine N in forma generale:

$$\sum_{k=0}^N a_k y(n-k) = \sum_{r=0}^M b_r x(n-r) \quad (2.53)$$

Applicando la trasformata z ad ambo i membri dell'eq. (2.53), otteniamo

$$\mathcal{Z}\left[\sum_{k=0}^N a_k y(n-k)\right] = \mathcal{Z}\left[\sum_{r=0}^M b_r x(n-r)\right]$$

che, in base alla proprietà 3 di tab. 2.1, possiamo riscrivere come

$$\sum_{k=0}^N a_k \mathcal{Z}[y(n-k)] = \sum_{r=0}^M b_r \mathcal{Z}[x(n-r)]$$

Se $X(z)$ e $Y(z)$ indicano rispettivamente le trasformate z di $x(n)$ e $y(n)$, dalla proprietà 4 di tab. 2.1 segue che

$$\mathcal{Z}[y(n-k)] = z^{-k} Y(z)$$

e

$$\mathcal{Z}[x(n-r)] = z^{-r} X(z)$$

Quindi

$$\sum_{k=0}^N a_k z^{-k} Y(z) = \sum_{r=0}^M b_r z^{-r} X(z)$$

Per la (2.52), è $H(z) = Y(z)/X(z)$, per cui

$$H(z) = \frac{\sum_{r=0}^M b_r z^{-r}}{\sum_{k=0}^N a_k z^{-k}} \quad (2.54)$$

La (2.54) esprime la funzione di trasferimento come funzione di z , e notiamo in particolare che i coefficienti dei polinomi al numeratore e al denominatore corrispondono, rispettivamente, ai coefficienti del secondo e del primo membro dell'equazione alle differenze (2.53).

Poiché la (2.54) è un rapporto di polinomi in z^{-1} , può anche essere espressa in forma fattorizzata³ come

$$H(z) = \frac{A \prod_{r=1}^M (1 - c_r z^{-1})}{\prod_{k=1}^N (1 - d_k z^{-1})} \quad (2.55)$$

Ciascuno dei fattori $(1 - c_r z^{-1})$ del numeratore della (2.55) contribuisce con uno zero in $z = c_r$ ed un polo in $z = 0$. Analogamente, ogni fattore della forma $(1 - d_k z^{-1})$ al denominatore dà luogo a un polo in $z = d_k$ ed a uno zero nell'origine. Caratteristica dei sistemi descrivibili con equazioni lineari alle differenze a coefficienti costanti è che le loro funzioni di trasferimento sono rapporti di polinomi in z^{-1} . Di conseguenza, la funzione di trasferimento può essere descritta, a meno del fattore scala A nella (2.55), da un diagramma di zeri e poli nel piano z .

La (2.54) non indica la regione di convergenza della funzione di trasferimento. Ciò è in accordo con il fatto che, come visto nel cap. 1, l'equazione alle differenze non specifica univocamente la risposta all'impulso di un sistema lineare invariante alla traslazione. Per la funzione di trasferimento (2.54) vi sono molte scelte per la regione di convergenza coerenti con il vincolo che esse corrispondano a regioni anulari limitate da poli (senza però contenerli). Per un dato rapporto di polinomi, ogni possibile scelta per la regione di convergenza conduce a una risposta all'impulso diversa, ma tutte quante corrispondono alla stessa equazione alle differenze. Se si assume che il sistema sia stabile, occorre scegliere la regione anulare che comprende il cerchio unitario. Se assumiamo che il sistema sia causale, allora scegliamo come regione di convergenza l'esterno di una circonferenza che passa per il polo di $H(z)$ più lontano dall'origine. Se il sistema è anche stabile, tutti i poli cadono all'interno del cerchio unitario e la regione di convergenza includerà il cerchio unitario. Per questa ragione è

³ La (2.55) assume che b_0 e a_0 nella (2.54) non siano zero. In generale, se b_0, b_1, \dots, b_{M-1} e a_0, a_1, \dots, a_{N-1} sono tutti zero, allora la (2.55) va espressa come

$$H(z) = \frac{Bz^{-M_1} \prod_{r=1}^{M-M_1} (1 - c_r z^{-1})}{z^{-N_1} \prod_{k=1}^{N-N_1} (1 - d_k z^{-1})}$$

spesso opportuno, quando si descrive la funzione di trasferimento in termini di un diagramma di poli e zeri nel piano z , includere il cerchio unitario nella figura per chiarire se i poli cadono dentro o fuori il cerchio unitario.

ESEMPIO. Come semplice esempio, consideriamo un sistema causale caratterizzato dall'equazione alle differenze

$$y(n) = ay(n-1) + x(n)$$

La funzione di trasferimento è

$$H(z) = \frac{1}{1 - az^{-1}} \quad (2.56)$$

e, per l'assunzione di causalità, la regione di convergenza è $|z| > |a|$, da cui notiamo che la risposta all'impulso è

$$h(n) = a^n u(n)$$

Nel caso particolare che sia $N = 0$ nella (2.54) o (2.53), il sistema non ha poli eccetto in $z = 0$ ed ha una risposta all'impulso di durata finita. Quando N è maggiore di zero, il sistema ha poli, ciascuno dei quali contribuisce con una sequenza esponenziale alla risposta al campione unitario. Perciò, se la funzione di trasferimento ha poli, la risposta all'impulso è di durata infinita.

Uno dei vantaggi della rappresentazione della funzione di trasferimento in termini di poli e zeri è che conduce a un metodo geometrico utile per ricavare la risposta in frequenza del sistema. Si è già visto in precedenza che la risposta del sistema a un'eccitazione sinusoidale può essere descritta in termini di risposta in frequenza, cioè del comportamento della funzione di trasferimento sul cerchio unitario. In particolare, la risposta è sinusoidale con la stessa frequenza dell'ingresso, e l'ampiezza dell'uscita è uguale all'ampiezza dell'ingresso moltiplicata per il valore assoluto della funzione di trasferimento alla frequenza di eccitazione. Lo sfasamento dell'uscita è uguale alla fase del numero complesso che rappresenta la funzione di trasferimento alla frequenza di eccitazione. Per valutare la funzione di trasferimento sul cerchio unitario, in corrispondenza di una frequenza di eccitazione ω_0 , occorre sostituire $z = e^{j\omega_0}$ nella (2.55). Consideriamo, per esempio, un fattore $(1 - c_r z^{-1})$, cui sono associati lo zero e il polo mostrati in fig. 2.7. In $z = e^{j\omega_0}$ il modulo del numero complesso rappresentato da questo fattore è uguale alla lunghezza del vettore che unisce lo zero con il punto corrispondente a ω_0 sul cerchio unitario, diviso per la lunghezza del vettore che congiunge il polo con lo stesso punto sul cerchio unitario. La fase del numero complesso è uguale alla fase del vettore associato allo zero meno la fase del vettore associato al polo. Il modulo del

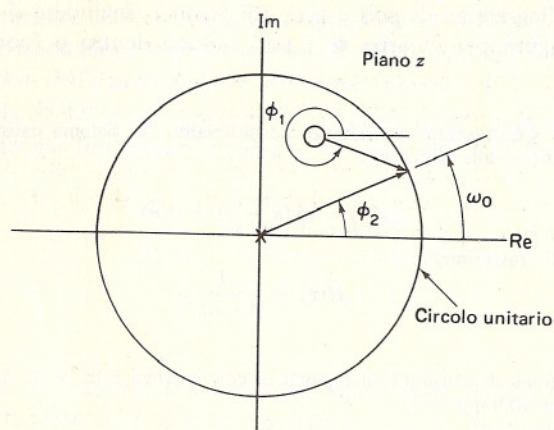


Fig. 2.7 Determinazione della risposta in frequenza dalla configurazione di poli e zeri con il metodo geometrico.

numero complesso rappresentato da un prodotto di fattori di questo tipo è il prodotto dei moduli, e poiché la fase del numero complesso prodotto è la somma delle fasi, la risposta in frequenza complessiva può essere ottenuta tramite l'effetto risultante dei vettori associati agli zeri e ai poli. Per esempio, in fig. 2.8 è mostrata la configurazione di poli e zeri per un

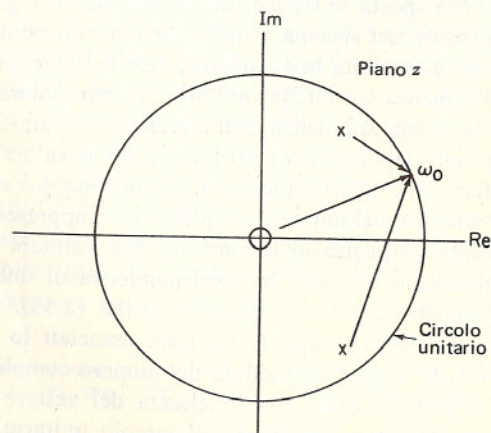


Fig. 2.8 Determinazione per via geometrica della risposta in frequenza di un sistema del secondo ordine.

sistema del secondo ordine. È chiaro dalla disposizione dei vettori dei poli e degli zeri che il modulo della risposta in frequenza ha un picco in prossimità dei poli. Da questa rappresentazione geometrica dovrebbe anche risultare chiaro che poli o zeri nell'origine non danno contributi al modulo della risposta in frequenza e introducono solo una componente lineare nella fase. Questi concetti sono illustrati in fig. 2.9, che mostra il diagramma di poli e zeri e la risposta in frequenza per un'equazione alle differenze del primo ordine, corrispondente alla funzione di trasferimento $H(z) = 1/(1 - az^{-1})$ e alla risposta all'impulso $h(n) = a^n u(n)$.

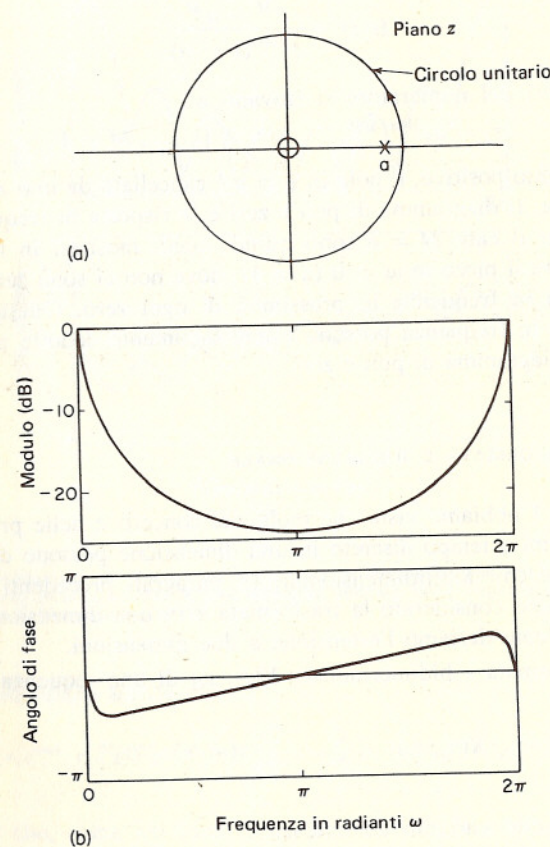


Fig. 2.9 Configurazione di poli e zeri per un filtro del primo ordine e risposta in frequenza corrispondente.

Come secondo esempio, sia la risposta all'impulso una versione troncata della risposta all'impulso dell'esempio precedente, cioè

$$h(n) = \begin{cases} a^n, & 0 \leq n \leq M-1 \\ 0, & \text{altrove} \end{cases}$$

Allora la funzione di trasferimento è

$$H(z) = \sum_{n=0}^{M-1} a^n z^{-n} = \frac{1 - a^M z^{-M}}{1 - a z^{-1}} \quad (2.57)$$

che può essere scritta

$$H(z) = \frac{z^M - a^M}{z^{M-1}(z - a)}$$

Poiché gli zeri del numeratore si trovano a

$$z_k = a e^{j(2\pi/M)k}, \quad k = 0, 1, \dots, M-1$$

dove a è assunto positivo, il polo in $z = a$ è cancellato da uno zero in posizione identica. Il diagramma di poli e zeri e la risposta in frequenza corrispondente per il caso $M = 8$ sono quindi quelli mostrati in fig. 2.10. Si possono notare il picco in $\omega = 0$ ($z = 1$), dove non ci sono zeri, e le valli nella risposta in frequenza in prossimità di ogni zero. Queste proprietà della risposta in frequenza possono essere facilmente dedotte per via geometrica dal diagramma di poli e zeri.

2.5. LA TRASFORMATA z BIDIMENSIONALE

Nel cap. 1 abbiamo visto che molti dei concetti e delle proprietà dei segnali e sistemi a tempo discreto in una dimensione possono essere estesi a segnali e sistemi multidimensionali. In paragrafi precedenti di questo capitolo abbiamo considerato la trasformata z in una dimensione. In questo paragrafo consideriamo l'estensione a due dimensioni.

La trasformata z bidimensionale $X(z_1, z_2)$ di una sequenza $x(m, n)$ è definita come

$$X(z_1, z_2) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(m, n) z_1^{-m} z_2^{-n} \quad (2.58)$$

dove z_1 e z_2 sono variabili complesse.

Esprimendo z_1 e z_2 in forma polare come

$$z_1 = r_1 e^{j\omega_1}$$

$$z_2 = r_2 e^{j\omega_2}$$

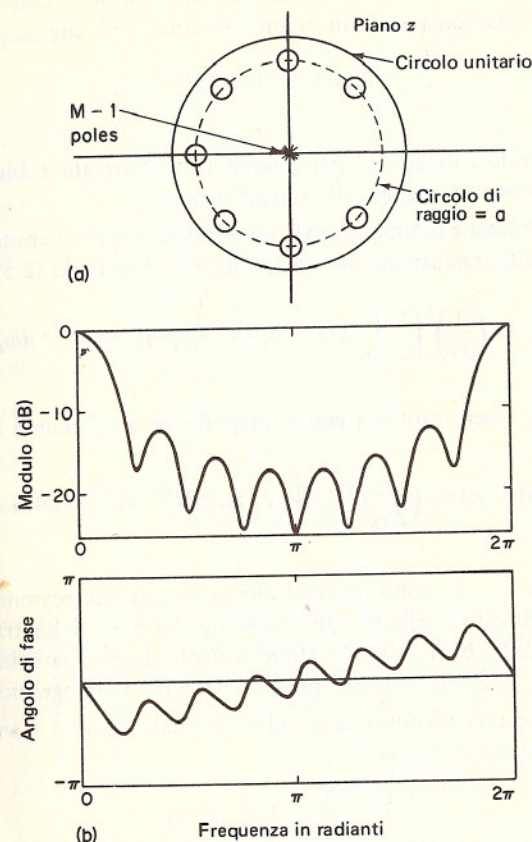


Fig. 2.10 Configurazione di poli e zeri e risposta in frequenza per un sistema FIR con risposta all'impulso che è una versione troncata della risposta all'impulso dello esempio illustrato in fig. 2.9.

la (2.58) può essere scritta come

$$X(r_1 e^{j\omega_1}, r_2 e^{j\omega_2}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(m, n) r_1^{-m} r_2^{-n} e^{-j\omega_1 m} e^{-j\omega_2 n} \quad (2.59)$$

Osserviamo che, come nel caso monodimensionale, la trasformata z bidimensionale può essere interpretata come la trasformata di Fourier bidimensionale della sequenza ottenuta moltiplicando $x(m, n)$ per la sequenza esponenziale bidimensionale $r_1^{-m} r_2^{-n}$. Per $|z_1| = |z_2| = 1$, cioè per $r_1 = r_2 = 1$, la trasformata z è uguale alla trasformata di Fourier. Per la con-

vergenza della trasformata z bidimensionale occorre che la sequenza $x(m, n)z_1^{-m}z_2^{-n}$ sia sommabile in valore assoluto, cioè che

$$\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} |x(m, n)z_1^{-m}z_2^{-n}| < \infty \quad (2.60)$$

L'insieme di valori di z_1 e z_2 per i quali la trasformata z bidimensionale converge definisce la regione di convergenza.

La trasformata z bidimensionale inversa può essere ottenuta applicando la relazione della trasformata di Fourier inversa (1.44) alla (2.59) ottenendo

$$x(m, n) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} X(r_1 e^{j\omega_1}, r_2 e^{j\omega_2}) r_1^m r_2^n e^{j\omega_1 m} e^{j\omega_2 n} d\omega_1 d\omega_2 \quad (2.61)$$

che può anche essere espressa come integrale su un percorso chiuso nella forma:

$$x(m, n) = \left(\frac{1}{2\pi j}\right)^2 \oint_{C_1} \oint_{C_2} X(z_1, z_2) z_1^{m-1} z_2^{n-1} dz_1 dz_2 \quad (2.62)$$

dove i percorsi C_1 e C_2 sono percorsi chiusi interni alla regione di convergenza e racchiudono l'origine. Diversamente dal caso della trasformata z monodimensionale, è in generale molto difficile determinare la regione di convergenza di $X(z_1, z_2)$ e quindi anche i percorsi di integrazione C_1 e C_2 .

Si dice che una trasformata z bidimensionale $X(z_1, z_2)$ è separabile se può essere espressa in forma

$$X(z_1, z_2) = X_1(z_1)X_2(z_2)$$

$X(z_1, z_2)$ sarà separabile solo se la sequenza $x(m, n)$ da cui essa è derivata è separabile, cioè se $x(m, n) = x_1(m)x_2(n)$. In tal caso $X_1(z_1)$ e $X_2(z_2)$ sono rispettivamente le trasformate z monodimensionali di $x_1(m)$ e $x_2(n)$.

Tutte le proprietà delle trasformate z monodimensionali riassunte nella tab. 2.1 si estendono facilmente a due dimensioni come indicato nella tab. 2.2, e le loro dimostrazioni sono analoghe a quelle del caso monodimensionale.

La trasformata z bidimensionale di una convoluzione di due sequenze bidimensionali è il prodotto delle loro trasformate z . Ne segue che la relazione ingresso-uscita per un sistema bidimensionale lineare invariante alla traslazione, espressa in termini di trasformata z , corrisponde alla moltiplicazione delle trasformate z dell'ingresso e della risposta all'impulso. Come in una dimensione, la trasformata z della risposta all'impulso viene chiamata funzione di trasferimento. La funzione di trasferimento per $|z_1| = |z_2| = 1$ è la risposta in frequenza del sistema.

Tab. 2.2

Sequenza	Trasformata z
1. $x(m, n)$	$X(z_1, z_2)$
2. $y(m, n)$	$Y(z_1, z_2)$
3. $ax(m, n) + by(m, n)$	$aX(z_1, z_2) + bY(z_1, z_2)$
4. $x(m + m_0, n + n_0)$	$z_1^{m_0} z_2^{n_0} X(z_1, z_2)$
5. $a^m b^n x(m, n)$	$X(a^{-1}z_1, b^{-1}z_2)$
6. $mnx(m, n)$	$z_1 z_2 \frac{\partial^2 X(z_1, z_2)}{\partial z_1 \partial z_2}$
7. $x^*(m, n)$	$X^*(z_1^*, z_2^*)$
8. $x(-m, -n)$	$X(z_1^{-1}, z_2^{-1})$
9. $\text{Re}[x(m, n)]$	$\frac{1}{2}[X(z_1, z_2) + X^*(z_1^*, z_2^*)]$
10. $\text{Im}[x(m, n)]$	$\frac{1}{2j}[X(z_1, z_2) - X^*(z_1^*, z_2^*)]$
11. $x(m, n) * y(m, n)$	$X(z_1, z_2)Y(z_1, z_2)$
12. $x(m, n)y(m, n)$	$\left(\frac{1}{2\pi j}\right)^2 \oint_{C_1} \oint_{C_2} X\left(\frac{z_1}{v_1}, \frac{z_2}{v_2}\right) Y(v_1, v_2) v_1^{-1} v_2^{-1} dv_1 dv_2$

Quando il sistema può essere descritto per mezzo di un'equazione lineare alle differenze a coefficienti costanti, la funzione di trasferimento del sistema è un rapporto di polinomi bidimensionali. In particolare, se uscita e ingresso soddisfano l'equazione alle differenze

$$\sum_{k=0}^{M_1} \sum_{r=0}^{N_1} a_{kr} y(m-k, n-r) = \sum_{k=0}^{M_2} \sum_{r=0}^{N_2} b_{kr} x(m-k, n-r) \quad (2.63)$$

e se si prende la trasformata z di entrambi i membri della (2.63), allora, usando le proprietà della tab. 2.2, si ha che

$$Y(z_1, z_2) \left[\sum_{k=0}^{M_1} \sum_{r=0}^{N_1} a_{kr} z_1^{-k} z_2^{-r} \right] = X(z_1, z_2) \left[\sum_{k=0}^{M_2} \sum_{r=0}^{N_2} b_{kr} z_1^{-k} z_2^{-r} \right]$$

cosicché la funzione di trasferimento $H(z_1, z_2)$ è data da

$$H(z_1, z_2) = \frac{Y(z_1, z_2)}{X(z_1, z_2)} = \frac{\sum_{k=0}^{M_2} \sum_{r=0}^{N_2} b_{kr} z_1^{-k} z_2^{-r}}{\sum_{k=0}^{M_1} \sum_{r=0}^{N_1} a_{kr} z_1^{-k} z_2^{-r}} \quad (2.64)$$

Nel caso monodimensionale, quando la funzione di trasferimento consisteva in un rapporto di polinomi, essa poteva essere descritta in termini dei suoi poli e zeri, cioè delle radici dei polinomi numeratore e denomi-

natore. Invece, un polinomio generico bidimensionale non può, in generale, essere scomposto in fattori. Ciò costituisce una rilevante differenza tra i casi monodimensionale e bidimensionale.

Confrontando la (1.38), che stabilisce le condizioni per la stabilità di un sistema bidimensionale invariante alla traslazione, con la (2.60), che esprime il requisito per la convergenza della trasformata z bidimensionale, vediamo che un sistema è stabile se e solo se la sua funzione di trasferimento converge per $|z_1| = |z_2| = 1$. Per sistemi causali monodimensionali la stabilità si verificava facilmente esaminando le posizioni dei poli. Per tale caso condizione necessaria e sufficiente per la stabilità è che tutti i poli della funzione di trasferimento giacciono all'interno del cerchio unitario. Nel caso speciale in cui la risposta all'impulso o, in modo equivalente, la funzione di trasferimento del sistema sia separabile, la stessa condizione si può applicare per esaminare la stabilità di un filtro causale. Ciò deriva dal fatto che per il caso separabile, con $h(m,n) = h_1(m)h_2(n)$, la risposta all'impulso sarà sommabile in valore assoluto se e solo se $h_1(m)$ e $h_2(n)$ sono sommabili in valore assoluto. Perciò un sistema causale separabile sarà stabile se e solo se i poli di $H_1(z_1)$ giacciono all'interno del cerchio unitario nel piano z_1 e i poli di $H_2(z_2)$ giacciono all'interno del cerchio unitario nel piano z_2 .

Nel caso generale, condizione necessaria e sufficiente perchè un rapporto di polinomi bidimensionali corrisponda a una funzione di trasferimento causale e stabile è che il polinomio denominatore non sia zero se $|z_1|$ e $|z_2|$ sono entrambi maggiori di uno [4,5]. Si può vedere che ciò è in accordo con la condizione discussa in precedenza per un sistema separabile, dal momento che, ad esempio, un polo di $H_1(z_1)$ all'esterno del cerchio unitario, cioè per $|z_1| > 1$, comporterà che il polinomio denominatore della funzioni di trasferimento bidimensionale sia zero per tale valore di z_1 qualunque sia il valore di z_2 e quindi inclusi i valori di z_2 per cui $|z_2| > 1$.

Un modo di applicare il teorema della stabilità appena visto è quello di scomporre in fattori il polinomio denominatore come polinomio in z_1 , così che le radici risultano funzioni di z_2 . In questo caso i poli di $H(z_1, z_2)$ sono funzioni di z_2 e si richiede che questi poli possano trovarsi soltanto all'esterno del cerchio unitario nel piano z_1 per valori di z_2 per cui $|z_2| < 1$.

ESEMPIO. Si consideri una $H(z_1, z_2)$ della forma

$$H(z_1, z_2) = \frac{1}{1 - z_1^{-1} + 2z_2^{-1} - z_1^{-1}z_2^{-1}}$$

o

$$H(z_1, z_2) = \frac{1}{(1 + 2z_2^{-1}) - z_1^{-1}(1 + z_2^{-1})} \quad (2.65)$$



Allora le radici del polinomio denominatore sono date da

$$z_1 = \frac{1 + z_2^{-1}}{1 + 2z_2^{-1}} \quad (2.66)$$

Per la stabilità si richiede che nella (2.66), per tutti i valori di z_2 per cui $|z_2| \geq 1$, il modulo di z_1 sia minore dell'unità, cioè richiediamo che per $|z_2| \geq 1$, sia

$$\left| \frac{1 + 2z_2^{-1}}{1 + z_2^{-1}} \right| < 1$$

$$|1 + z_2^{-1}| < |1 + 2z_2^{-1}|$$

In maniera equivalente, ciò può essere scritto come

$$|1 + z_2^{-1}|^2 < |1 + 2z_2^{-1}|^2$$

E immediato verificare che per $z_2^{-1} = -\frac{1}{3} + j\frac{1}{3}$ questa disuguaglianza non è soddisfatta, e di conseguenza la funzione di trasferimento (2.65) non corrisponde ad un sistema stabile.

È chiaro che, mentre la condizione di stabilità sopra discussa è immediata in teoria, in pratica è difficile da applicare. Ci sono diverse altre formulazioni delle condizioni di stabilità, che non trattiamo, che sono meno immediate dal punto di vista teorico ma sono forse leggermente più semplici da applicare.

SOMMARIO

In questo capitolo abbiamo generalizzato molti degli argomenti discussi nel cap. 1. In particolare, abbiamo presentato la definizione della trasformata z e la regione di convergenza associata a sequenze monolateri destre, sinistre e bilateri. Si è quindi presa in esame la trasformata z inversa insieme ai metodi per calcolarla (integrazione complessa, espansione in fratti semplici e uso delle serie di potenze). Sono state anche presentate diverse proprietà della trasformata z simili a quelle viste nel cap. 1 per la trasformata di Fourier.

La rappresentazione di sistemi lineari invarianti alla traslazione in termini della trasformata z ci ha portati a trattare la funzione di trasferimento. Per sistemi caratterizzati da equazioni lineari alle differenze a coefficienti costanti, la funzione di trasferimento è un rapporto di polinomi, e può quindi essere caratterizzata in termini di una configurazione di poli e zeri nel piano z . Questa rappresentazione porta, fra l'altro, a un utile metodo geometrico per ottenere la risposta in frequenza del sistema.

Il capitolo si è concluso con una breve introduzione alla trasformata z bidimensionale.



BIBLIOGRAFIA

1. J. R. Ragazzini and G. F. Franklin, *Sampled Data Control Systems*, McGraw-Hill Book Company, New York, 1958.
2. E. I. Jury, *Theory and Application of the z-Transform Method*, John Wiley & Sons, Inc., New York, 1964.
3. R. V. Churchill, *Complex Variables and Applications*, McGraw-Hill Book Company, New York, 1960.
4. J. L. Shanks, S. Treitel, and J. H. Justice, "Stability and Synthesis of Two-Dimensional Recursive Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 3, June 1972, pp. 115-128.
5. T. S. Huang, "Stability of Two-Dimensional Recursive Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 2, June 1972, pp. 158-163.

PROBLEMI

1. Per ognuna delle seguenti sequenze determinare la rispettiva trasformata z , indicando anche la regione di convergenza:

- (a) $(\frac{1}{2})^n u(n)$.
- (b) $-(\frac{1}{2})^n u(-n-1)$.
- (c) $(\frac{1}{2})^n u(-n)$.
- (d) $\delta(n)$.
- (e) $\delta(n-1)$.
- (f) $\delta(n+1)$.
- (g) $(\frac{1}{2})^n [u(n) - u(n-10)]$.

2. Determinare la trasformata z di ognuna delle seguenti sequenze. Completare la risposta con la regione di convergenza nel piano z e con uno schizzo della disposizione dei poli e degli zeri. Esprimere tutte le sommatorie in forma chiusa (α può essere complesso).

- (a) $x(n) = \alpha^n$, $0 < |\alpha| < 1$.
- (b) $x(n) = Ar^n \cos(\omega_0 n + \phi)u(n)$, $0 < r < 1$.
- (c) $x(n) = \begin{cases} 1, & 0 \leq n \leq N-1, \\ 0, & N \leq n, \\ 0, & n < 0. \end{cases}$
- (d) $x(n) = \begin{cases} n, & 0 \leq n \leq N, \\ 2N-n, & N+1 \leq n \leq 2N, \\ 0, & 2N \leq n, \\ 0, & 0 > n. \end{cases}$

[Suggerimento: per prima cosa esprimere $x(n)$ in termini della $x(n)$ del punto (c).]

3. Si consideri la sequenza $y(n)$ data da $y(n) = x(n) * h(n)$, dove $h(n) = (1+j)^n u(n)$ e $|x(n)| \leq 1$. La sequenza $|y(n)|$ è necessariamente limitata?

4. (a) Sono qui elencate diverse trasformate z . Per ognuna di esse calcolare la trasformata z inversa usando tutti e tre i metodi (integrale di linea, espansione in fratti semplici, divisione lunga) descritti nel par. 2.2.

$$X(z) = \frac{1}{1 + \frac{1}{2}z^{-1}}, \quad |z| > \frac{1}{2}$$

$$X(z) = \frac{1}{1 + \frac{1}{2}z^{-1}}, \quad |z| < \frac{1}{2}$$

$$X(z) = \frac{1 - \frac{1}{2}z^{-1}}{1 + \frac{3}{4}z^{-1} + \frac{1}{8}z^{-2}}, \quad |z| > \frac{1}{2}$$

$$X(z) = \frac{1 - \frac{1}{2}z^{-1}}{1 - \frac{1}{4}z^{-2}}, \quad |z| > \frac{1}{2}$$

$$X(z) = \frac{1 - az^{-1}}{z^{-1} - a}, \quad |z| > |1/a|$$

- (b) Si consideri la sequenza monolatera destra $x(n)$ con trasformata z

$$X(z) = \frac{1}{(1 - az^{-1})(1 - bz^{-1})} = \frac{z^2}{(z - a)(z - b)}$$

Nel par. 2.2 abbiamo considerato il calcolo di $x(n)$ per mezzo dell'espansione in fratti semplici di $X(z)$ considerata come un rapporto di polinomi in z^{-1} . Eseguire un'espansione in fratti semplici di $X(z)$ considerata come rapporto di polinomi in z e da questa espansione ricavare $x(n)$.

5. Nel par. 2.2 abbiamo osservato che per $n < 0$ è spesso più conveniente calcolare la trasformata inversa (2.22) tramite la sostituzione di variabili $z = p^{-1}$, ottenendo così l'espressione (2.28). Se si integra lungo il cerchio unitario, ad esempio, ciò ha l'effetto di mappare l'interno del cerchio unitario nell'esterno e viceversa. È chiaro, comunque, che senza questa sostituzione possiamo ottenere $x(n)$, per $n < 0$, calcolando i residui nei poli multipli in $z = 0$.

Sia

$$X(z) = \frac{1}{1 - \frac{1}{2}z}$$

dove la regione di convergenza comprende il cerchio unitario.

- (a) Ricavare $x(0)$, $x(-1)$ e $x(-2)$ mediante il calcolo esplicito della (2.22), valutando cioè il residuo nei poli in $z = 0$.
 - (b) Ricavare $x(n)$ per $n > 0$ mediante la (2.22) e per $n < 0$ mediante la (2.28).
6. Determinare una sequenza $x(n)$ la cui trasformata z sia $X(z) = e^z + e^{1/z}$, $z \neq 0$.
 7. Determinare se la funzione $F(z) = z^*$ può corrispondere alla trasformata z di una sequenza. Motivare la risposta.
 8. Supponiamo che $F(z)$ sia una funzione razionale, cioè

$$F(z) = \frac{N(z)}{D(z)}$$

dove $N(z)$ e $D(z)$ sono polinomi. Assumiamo inoltre che $F(z)$ non abbia poli o zeri di molteplicità maggiore di uno. Sia C una curva chiusa semplice, e Z e P , rispettivamente il numero di zeri e poli di $F(z)$ interni alla curva (si assuma che non ci siano poli o zeri su C).

(a) Dimostrare che

$$\frac{1}{2\pi j} \oint_C \frac{F'(z)}{F(z)} dz = Z - P$$

dove $F'(z)$ è la derivata di $F(z)$.

(b) Esprimendo $F(z)$ in forma polare come $F(z) = |F(z)|e^{j\arg F(z)}$, dimostrare che il cambiamento di $\arg[F(z)]$ quando C è percorso esattamente una volta è $2\pi(Z - P)$.

(Si può dimostrare che questo risultato è generalizzabile al caso di poli e zeri multipli se si contano poli e zeri secondo la loro molteplicità, vale a dire un polo di second'ordine è contato due volte).

9. Si indichi con $X(z)$ un rapporto di polinomi in z , cioè

$$X(z) = \frac{P(z)}{Q(z)}$$

Dimostrare che se $X(z)$ ha un polo del primo ordine in $z = z_0$, allora

$$\text{Res}[X(z) \text{ at } z = z_0] = \frac{P(z_0)}{Q'(z_0)}$$

dove $Q'(z_0)$ indica la derivata di $Q(z)$ calcolata in $z = z_0$.

10. Dimostrare che, se $X(z)$ è la trasformata z di $x(n)$, allora

(a) $z^n X(z)$ è la trasformata z di $x(n + n_0)$.

(b) $X(a^{-1}z)$ è la trasformata z di $a^n x(n)$.

(c) $-zX'(z)$ è la trasformata z di $nx(n)$.

11. Dimostrare che

$$\Im[x^*(n)] = X^*(z^*)$$

$$\Im[x(-n)] = X\left(\frac{1}{z}\right)$$

$$\Im[\text{Re } x(n)] = \frac{1}{2}[X(z) + X^*(z^*)]$$

$$\Im[\text{Im } x(n)] = \frac{1}{2j}[X(z) - X^*(z^*)]$$

dove $X(z)$ indica la trasformata z di $x(n)$.

12. Determinare la trasformata z di $n^2 x(n)$ espressa in termini della trasformata z di $x(n)$.

13. La sequenza di autocorrelazione $c(n)$ di una sequenza $x(n)$ è definita come

$$c(n) = \sum_{k=-\infty}^{\infty} x(k)x(n+k)$$

Determinare la trasformata z di $c(n)$ in termini della trasformata z di $x(n)$.

14. Si indichi con $x(n)$ una sequenza causale, cioè tale che $x(n) = 0$ per $n < 0$. Inoltre si assuma che $x(0) \neq 0$.

(a) Dimostrare che non ci sono poli o zeri di $X(z)$ per $z = \infty$.

(b) Dimostrare che nel piano z finito il numero dei poli è uguale al numero degli zeri (il piano z finito esclude $z = \infty$).

15. Si consideri un filtro con risposta all'impulso finita, la cui risposta all'impulso $h(n)$ abbia lunghezza $(2N + 1)$. Se $h(n)$ è reale e pari, dimostrare che gli zeri della funzione di trasferimento si verificano a coppie con simmetria speculare rispetto al circolo unitario, cioè, se $H(z) = 0$ per $z = \rho e^{j\theta}$, allora $H(z) = 0$ anche per $z = (1/\rho)e^{j\theta}$.

16. Per una sequenza $x(n)$, nulla per $n < 0$, dimostrare che $\lim_{z \rightarrow \infty} X(z) = x(0)$. Qual è il teorema corrispondente per una sequenza nulla per $n > 0$?

17. Si consideri una sequenza $x(n)$ la cui trasformata z sia

$$X(z) = \frac{\frac{1}{3}}{1 - \frac{1}{2}z^{-1}} + \frac{\frac{1}{4}}{1 - 2z^{-1}}$$

e per cui la regione di convergenza include il circolo unitario. Si usino i teoremi del precedente probl. 16 per calcolare $x(0)$.

18. I poli e gli zeri della trasformata z di una sequenza reale $x(n)$ sono tutti all'interno del circolo unitario. Determinare in termini di $x(n)$ una sequenza reale $x_1(n)$, diversa da $x(n)$, ma tale per cui $x_1(0) = x(0)$, $|x_1(n)| = |x(n)|$ e la trasformata z di $x_1(n)$ ha poli e zeri tutti interni al circolo unitario.

19. Sia $x(n)$ una sequenza di durata finita di lunghezza N con $x(n) = 0$ per $n < 0$ e $x(n) = 0$ per $n \geq N$. Non si assuma che $x(n)$ è reale. Dal seguente elenco scegliere il valore che corrisponde al numero di possibili scelte distinte per $x(n)$ nel caso sia specificato il modulo della sua trasformata di Fourier.

- | | | |
|----------------|-------------------|---------------------|
| (a) 1 | (f) N^2 | (j) $\log_2(N - 1)$ |
| (b) N | (g) $N(N - 1)$ | (k) N^N |
| (c) $\log_2 N$ | (h) 2^{N-1} | (l) $N!$ |
| (d) 2^N | (i) $N \cdot 2^N$ | (m) $(N - 1)!$ |
| (e) ∞ | | |

20. Si consideri un insieme di N sequenze distinte

$$S = \{x_v(n)\}, \quad v = 0, 1, \dots, N - 1$$

e il corrispondente insieme delle trasformate z

$$T = \{X_v(z)\}$$

dove $X_v(z)$ è la trasformata z di $x_v(n)$. Gli elementi di S e T hanno le seguenti proprietà:

- (1) Le x_v sono reali.
 - (2) Le x_v sono causali, nel senso che $x_v = 0$ per $n < 0$.
 - (3) Le x_v sono stabili; cioè la regione di convergenza di $X_v(z)$ contiene il circolo unitario.
 - (4) Le X_v sono funzioni razionali di z ; vale a dire che esse possono essere espresse come rapporto di polinomi in z (o z^{-1}).
 - (5) I moduli di tutte le X_v sono uguali sul circolo unitario; cioè, $|X_\mu(e^{j\omega})| = |X_\nu(e^{j\omega})|$ per tutti i ν e μ e per $-\pi < \omega < \pi$.
 - (6) S e T sono completi; in altre parole, se $X_\nu \in T$ e $|X_\mu(e^{j\omega})| = |X_\nu(e^{j\omega})|$ per $-\pi < \omega < \pi$, allora $X_\mu \in T$.
- (a) Dimostrare che $X_\nu(z)$ può essere espressa come

$$X_\mu(z) = R_{\mu\nu}(z)X_\nu(z)$$

dove $R_{\mu\nu}(z)$ è una funzione razionale di z (o z^{-1}) con modulo unitario sul circolo unitario, cioè

$$|R_{\mu\nu}(e^{j\omega})| = 1, \quad -\pi < \omega \leq \pi$$

$R_{\mu\nu}(z)$ è necessariamente stabile? Perché?

- (b) Indicare con $X_0(z)$ l'elemento di T che ha il minor numero di zeri esterni al circolo unitario. Per mezzo del teorema del valore iniziale, dimostrare che l'elemento di S che ha modulo massimo nell'origine è $x_0(n)$, cioè

$$|x_0(0)| > |x_v(0)|, \quad \text{per tutti } v \neq 0$$

- (c) Si supponga che la funzione di trasferimento $H(z)$ di un filtro stabile e causale sia una funzione razionale di z . Sia inoltre, sul circolo unitario,

$$|H(e^{j\omega})| = |H(e^{-j\omega})|$$

e

$$\arg [H(e^{j\omega})] = -\arg [H(e^{-j\omega})]$$

$$-\pi < \omega \leq \pi$$

Si supponga però che gli zeri di $H(z)$ non siano tutti interni al circolo unitario. Come è possibile ricavare da $H(z)$ un filtro $G(z)$, stabile e causale, con risposta all'impulso reale, che abbia tutti gli zeri interni al circolo unitario e per cui valga

$$|G(e^{j\omega})| = |H(e^{j\omega})|, \quad -\pi < \omega \leq \pi?$$

[Suggerimento: utilizzare il risultato del punto (a).]

21. Si supponga che sia data una sequenza $x(n)$ la cui trasformata di Fourier gode della seguente proprietà:

$$\begin{aligned} X(e^{j\omega}) &\neq 0, & |\omega| < \omega_c \\ &= 0, & \omega_c < |\omega| < \pi \end{aligned}$$

- (a) Dimostrare che se si definisce una nuova sequenza $x_1(n)$ di valori

$$x_1(n) = x(Mn), \quad n = 0, \pm 1, \pm 2, \dots$$

(cioè si prende un campione ogni M), allora è

$$X_1(z) = \frac{1}{M} \sum_{l=0}^{M-1} X(z^{1/M} e^{-j(2\pi/M)l})$$

- (b) Disegnare la trasformata di Fourier di $x_1(n)$ per il caso $\omega_c = \pi/M$ [assumere una forma arbitraria per $X(e^{j\omega})$].

- (c) Si supponga ora di avere una sequenza $x_1(n)$ e di definire una nuova sequenza $x_2(n)$ di valori

$$x_2(n) = \begin{cases} x_1\left(\frac{n}{M}\right), & n = 0, \pm M, \pm 2M, \dots \\ 0, & \text{altrove} \end{cases}$$

Si dimostri che allora è

$$X_2(z) = X_1(z^M)$$

- (d) Disegnare la trasformata di Fourier di $x_2(n)$ quando $x_1(n)$ è quella del punto (b).

- (e) Utilizzando i risultati precedenti, dimostrare come la sequenza originale $x(n)$ possa essere riottenuta esattamente da $x_2(n)$. Qual è la relazione generale tra M e ω_c che assicura il ripristino di $x(n)$?

22. Si consideri un sistema lineare invariante alla traslazione con risposta all'impulso $h(n)$ e con ingresso $x(n)$ dato da

$$h(n) = \begin{cases} a^n, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

$$x(n) = \begin{cases} 1, & 0 \leq n \leq (N-1) \\ 0, & \text{altrove} \end{cases}$$

- (a) Determinare l'uscita $y(n)$ calcolando esplicitamente la convoluzione discreta di $x(n)$ e $h(n)$.
(b) Determinare l'uscita $y(n)$ calcolando la trasformata z inversa del prodotto delle trasformate z dell'ingresso e della risposta all'impulso.
23. Si consideri un sistema discreto, lineare e invariante alla traslazione, per cui l'ingresso $x(n)$ e l'uscita $y(n)$ sono legati dall'equazione alle differenze finite del primo ordine

$$y(n) + \frac{1}{2}y(n-1) = x(n)$$

Scegliere dal seguente elenco due possibili risposte all'impulso per tale sistema.

- | | | |
|-------------------------------|------------------------------|---|
| (a) $(-\frac{1}{2})^n u(n)$ | (e) $(\frac{1}{2})^n u(n-1)$ | (h) $(-\frac{1}{2})^n u(-n-1)$ |
| (b) $(2^n) u(n)$ | (f) $(-2)^n u(-n-1)$ | (i) $\frac{1}{2}(-\frac{1}{2})^{n-1} u(-n-1)$ |
| (c) $(n)^{1/2} u(n)$ | (g) $(\frac{1}{2})^n u(n)$ | (j) $2(-2)^{n-1} u(-n-1)$ |
| (d) $(\frac{1}{2})^{-n} u(n)$ | | |

24. Un sistema lineare invariante alla traslazione e causale è definito dall'equazione alle differenze

$$y(n) = y(n-1) + y(n-2) + x(n-1)$$

- (a) Trovare la funzione di trasferimento $H(z) = Y(z)/X(z)$ del sistema. Rappresentare nel piano z i poli e gli zeri di $H(z)$ e indicare la regione di convergenza.
(b) Trovare la risposta all'impulso del sistema.
(c) Dovreste aver trovato che questo sistema è instabile. Determinare una risposta all'impulso stabile (non causale) che soddisfi l'equazione alle differenze precedente.

25. Si consideri un sistema discreto, lineare e invariante alla traslazione, con ingresso $x(n)$ e uscita $y(n)$, per cui vale

$$y(n-1) - \frac{5}{2}y(n) + y(n+1) = x(n)$$

Il sistema può non essere stabile e/o causale.

In relazione alla configurazione dei poli e degli zeri associata a questa equazione alle differenze, si determinino tre possibili risposte all'impulso del sistema. Verificare che tutte le scelte soddisfano l'equazione alle differenze.

26. Si consideri un sistema lineare invariante alla traslazione e causale con funzione di trasferimento

$$H(z) = \frac{1 - a^{-1}z^{-1}}{1 - az^{-1}}$$

dove a è reale.

- (a) Per quale insieme di valori di a il sistema è stabile?
(b) Per $0 < a < 1$ disegnare il diagramma dei poli e degli zeri, indicando la regione di convergenza.
(c) Dimostrare graficamente nel piano z che questo sistema è un sistema passatutto, vale a dire che il modulo della risposta in frequenza è costante.

- (d) $H(z)$ deve essere messo in cascata con un sistema $\hat{H}(z)$ in modo che la funzione di trasferimento complessiva sia unitaria. Con $0 < a < 1$ e $\hat{H}(z)$ stabile, determinare la sua risposta all'impulso $\hat{h}(n)$.
27. Si consideri un sistema a tempo discreto, lineare e invariante alla traslazione, con ingresso $x(n)$ e uscita $y(n)$, per cui vale

$$y(n-1) - \frac{1}{3}y(n) + y(n+1) = x(n)$$

Il sistema è stabile. Determinare la risposta all'impulso.

28. Sono qui riportate quattro trasformate z . Determinare quali di esse possono essere considerate come funzione di trasferimento di un sistema lineare a tempo discreto non necessariamente stabile, ma per il quale la risposta all'impulso sia nulla per $n < 0$. Giustificare chiaramente la risposta.

- (a) $(1 - z^{-1})^2 / (1 - \frac{1}{2}z^{-1})$.
 (b) $(z - 1)^2 / (z - \frac{1}{2})$.
 (c) $(z - \frac{1}{4})^5 / (z - \frac{1}{2})^6$.
 (d) $(z - \frac{1}{4})^6 / (z - \frac{1}{2})^5$.

29. Per mezzo delle trasformate z , determinare la risposta del sistema discreto lineare, causale e invariante alla traslazione, caratterizzato dall'equazione alle differenze finite

$$y(n) - 2r \cos \theta y(n-1) + r^2 y(n-2) = x(n)$$

quando l'eccitazione è

$$x(n) = \alpha^n u(n)$$

30. Una sequenza $x(n)$ è l'uscita di un sistema lineare invariante alla traslazione il cui ingresso è $s(n)$. Tale sistema è descritto dall'equazione alle differenze finite

$$x(n) = s(n) - e^{-8\pi} s(n-8)$$

con $0 < \alpha$.

- (a) Calcolare la funzione di trasferimento

$$H_1(z) = \frac{X(z)}{S(z)}$$

e disegnare i suoi poli e zeri nel piano z , indicando anche la regione di convergenza.

- (b) Si voglia ricostruire $s(n)$ da $x(n)$ attraverso un sistema lineare invariante alla traslazione. Determinare la funzione di trasferimento

$$H_2(z) = \frac{Y(z)}{X(z)}$$

tale che $y(n) = s(n)$. Determinare tutte le possibili regioni di convergenza per $H_2(z)$ e, per ognuna, dire se il sistema è stabile e causale oppure no.

- (c) Trovare tutte le possibili risposte all'impulso $h_2(n)$ tali che

$$y(n) = h_2(n) * x(n) = s(n)$$

31. Nel probl. 1.32 si è discussa la rappresentazione di una funzione a tempo continuo per mezzo di una sequenza corrispondente ai coefficienti di una espansione in serie in termini di un dato insieme di funzioni base. È utile anche rappresentare in maniera simile una sequenza per mezzo di un'altra. Perciò si può considerare l'espansione di una sequenza $f(n)$ in termini di un insieme di sequenze $\phi_k(n)$ come

$$f(n) = \sum_{k=0}^{\infty} g_k \phi_k(n) \quad (\text{P2.31-1})$$

I coefficienti g_k di questa espansione possono essere considerati come una nuova sequenza che rappresenta $f(n)$. In questo problema si considereranno sequenze $f(n)$ nulle per $n < 0$ e solo una particolare scelta per le funzioni $\phi_k(n)$. Più precisamente, sceglieremo quelle sequenze $\phi_k(n)$ tali che le loro trasformate z siano date da

$$\Phi_k(z) = \sum_{n=0}^{\infty} \phi_k(n) z^{-n} = \left(\frac{1 - az^{-1}}{z^{-1} - a} \right)^{-k}, \quad k \geq 0 \quad (\text{P2.31-2})$$

con $|a| < 1$ e con una regione di convergenza di $\Phi_k(z)$ tale che $\phi_k(n)$ sia nulla per $n < 0$. Si assuma il coefficiente a reale.

- (a) Si indichi con $F(z)$ la trasformata z di $f(n)$ e con $G(w)$ la « trasformata w » di g_k , cioè

$$F(z) = \sum_{n=0}^{\infty} f(n) z^{-n}$$

$$G(w) = \sum_{k=0}^{\infty} g_k w^{-k}$$

$G(w)$ e $F(z)$ sono legate attraverso una sostituzione di variabili, cioè

$$F(z) = G[R(z)]$$

Si determini $R(z)$ e si dimostri che, per $z = e^{j\omega}$, w può essere espresso nella forma $w = e^{j\theta}$, dove ω e θ sono reali. Non è necessario calcolare esplicitamente θ in funzione di ω .

Ciò dimostra, quindi, che le trasformate di Fourier della sequenza originale $f(n)$ e della nuova sequenza g_k sono legate da una trasformazione dell'asse frequenza. Siccome θ è funzione non lineare di ω , trasformare la sequenza $f(n)$ nella sequenza g_k corrisponde ad una distorsione nonlineare dell'asse delle frequenze. La parte rimanente del problema riguarda i metodi di calcolo di g_k .

- (b) Si può dimostrare che le sequenze $\phi_k(n)$ corrispondenti alla (P2.31-2) soddisfano la relazione

$$\sum_{n=0}^{\infty} n \phi_k(n) \phi_r(n) = \begin{cases} 0, & k \neq r \\ k, & k = r \end{cases}$$

In conseguenza di ciò, si può ottenere g_k per $k > 0$ dalla relazione

$$g_k = \frac{1}{k} \sum_{n=0}^{\infty} n \phi_k(n) f(n) \quad (\text{P2.31-3})$$

Si assuma che $f(n)$ abbia durata finita, cioè che sia $f(n) = 0$ per $n < 0$ e per $n > (N-1)$. Secondo la (P2.31-3), g_k può essere ottenuto, per $k > 0$, filtrando $f(-n)$ con un filtro numerico lineare e invariante alla traslazione. Si ricavi la risposta all'impulso di tale filtro in termini di $\phi_k(n)$ e la funzione di trasferimento (cioè la trasformata z della risposta all'impulso) in termini di $\Phi_k(z)$. Si specifichi inoltre come si può ottenere g_k dall'uscita del filtro; occorre un filtro diverso per ogni valore di k .

La procedura per calcolare g_k testè ricavata può essere usata solo per $k > 0$ a causa del fattore $1/k$ nella (P2.31-3). Per calcolare g_0 si può osservare, a partire dalla (P2.31-1), che è

$$f(0) = \sum_{k=0}^{\infty} g_k \phi_k(0)$$

ovvero

$$g_0 = \frac{1}{\phi_0(0)} f(0) - \sum_{k=1}^{\infty} g_k \phi_k(0) \quad (\text{P2.31-4})$$

(c) (1) Si determini $\phi_k(0)$.

(2) Come nel punto (b), si assuma che $f(n)$ sia di durata finita. Dimostrare, usando la (P2.31-4), che g_0 può essere ottenuto filtrando $f(-n)$ con un filtro numerico lineare ed invariante alla traslazione, e specificare come si può ricavare g_0 dall'uscita del filtro. Si esprima la *risposta all'impulso* di tale filtro come sommatoria di termini che comprendono le $\phi_k(n)$.

(3) Dalla sommatoria del punto (2), determinare la *funzione di trasferimento* del filtro numerico in *forma chiusa*.

Suggerimento: il risultato è un semplice filtro del primo ordine.

(Nota: I risultati di questo problema sono discussi nel riferimento [8] della Bibliografia del cap. 1).

32. Come si è visto nel par. 2.5, una condizione necessaria e sufficiente per la stabilità della funzione di trasferimento di un sistema bidimensionale è che il polinomio denominatore non si annulli quando $|z_1|$ e $|z_2|$ sono contemporaneamente maggiori o uguali a uno.

Si consideri la classe dei filtri numerici bidimensionali del primo ordine la cui funzione di trasferimento sia esprimibile nella forma

$$H(z_1, z_2) = \frac{1}{1 - az_1^{-1} - bz_2^{-1}}$$

Dimostrare che per questa classe di filtri, una condizione necessaria e sufficiente per la stabilità è che sia

$$|a| + |b| < 1$$

3. LA TRASFORMATTA DI FOURIER DISCRETA

3.0 INTRODUZIONE

Nei capitoli 1 e 2 abbiamo discusso la rappresentazione di sequenze e sistemi lineari invarianti alla traslazione in termini della trasformata di Fourier e della trasformata z . Nel caso particolare in cui la sequenza da rappresentare è di durata finita, cioè ha soltanto un numero finito di valori non nulli, è possibile sviluppare una rappresentazione di Fourier alternativa, chiamata *trasformata di Fourier discreta* (DFT). Come vedremo in questo capitolo, la DFT è una rappresentazione di Fourier di una sequenza di lunghezza finita che è essa stessa una sequenza anziché una funzione continua, e corrisponde a campioni egualmente spazati in frequenza della trasformata di Fourier del segnale. Oltre alla sua importanza dal punto di vista teorico come rappresentazione di Fourier di sequenze, la DFT, grazie all'esistenza di un metodo efficiente per il suo calcolo (che verrà esposto dettagliatamente nel cap. 6), svolge un ruolo centrale nella realizzazione di vari algoritmi di elaborazione numerica dei segnali [1,2].

Per ricavare ed interpretare la rappresentazione in termini di DFT di una sequenza di durata finita si possono adottare diversi punti di vista. Noi abbiamo scelto di basarci sulla relazione che esiste tra sequenze di lunghezza finita e sequenze periodiche, e pertanto consideriamo in primo luogo la rappresentazione di sequenze periodiche mediante serie di Fourier. Applicheremo poi tale rappresentazione anche al caso di una sequenza di lunghezza finita, mediante l'artificio di associare a questa una sequenza periodica coincidente in ogni suo periodo con la sequenza di lunghezza finita. Come vedremo, la rappresentazione mediante serie di Fourier di questa sequenza periodica corrisponde alla DFT della sequenza di lunghezza finita.

3.1 RAPPRESENTAZIONE DI SEQUENZE PERIODICHE - LA SERIE DI FOURIER DISCRETA

Si consideri una sequenza $\tilde{x}(n)$ periodica¹ con periodo N , cioè tale che sia $\tilde{x}(n) = \tilde{x}(n + kN)$ per ogni valore intero di k . Tale sequenza non può essere rappresentata mediante la sua trasformata z , poichè non esiste alcun valore di z per cui la trasformata z converge. È possibile, tuttavia, rappresentare $\tilde{x}(n)$ per mezzo di una serie di Fourier, cioè come somma di sequenze sinusoidali o cosinusoidali o, in modo equivalente, di sequenze esponenziali complesse con frequenze multipli interi della frequenza fondamentale $2\pi/N$ associata alla sequenza periodica. A differenza della serie di Fourier valida per funzioni periodiche continue, esistono soltanto N esponenziali complessi distinti il cui periodo è un sottomultiplo intero del periodo fondamentale N . Ciò deriva dal fatto che l'esponenziale complesso

$$e_k(n) = e^{j(2\pi/N)nk} \quad (3.1)$$

è periodico in k con periodo N . Perciò $e_0(n) = e_N(n)$, $e_1(n) = e_{N+1}(n)$ ecc., e quindi l'insieme di N esponenziali complessi rappresentati nella (3.1) con $k = 0, 1, 2, \dots, N-1$ definisce tutti gli esponenziali complessi distinti con frequenze che sono multipli interi di $2\pi/N$. Pertanto, per la rappresentazione in serie di Fourier di una sequenza periodica, $\tilde{x}(n)$, bastano soltanto N di questi esponenziali complessi e quindi essa può scriversi nella forma

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) e^{j(2\pi/N)nk} \quad (3.2)$$

La costante moltiplicativa $1/N$ è stata inserita per convenienza e, naturalmente, non ha nessun effetto importante sulla natura della rappresentazione. Per ottenere i coefficienti $\tilde{X}(k)$ dalla sequenza periodica $\tilde{x}(n)$, usiamo il fatto che risulta

$$\frac{1}{N} \sum_{n=0}^{N-1} e^{j(2\pi/N)nr} = \begin{cases} 1, & \text{per } r = mN, m \text{ intero} \\ 0, & \text{altrove} \end{cases} \quad (3.3)$$

Perciò moltiplicando entrambi i membri della relazione (3.2) per $e^{-j(2\pi/N)nr}$ e sommando da $n = 0$ a $n = N-1$, otteniamo

$$\sum_{n=0}^{N-1} \tilde{x}(n) e^{-j(2\pi/N)nr} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{X}(k) e^{j(2\pi/N)(k-r)n}$$

¹ D'ora in poi useremo il simbolo \sim per indicare sequenze periodiche ogni volta che è importante distinguere chiaramente sequenze periodiche e aperiodiche.

Oppure, scambiando l'ordine delle sommatorie al secondo membro dell'espressione precedente, si può scrivere

$$\sum_{n=0}^{N-1} \tilde{x}(n) e^{-j(2\pi/N)nr} = \sum_{k=0}^{N-1} \tilde{X}(k) \left[\frac{1}{N} \sum_{n=0}^{N-1} e^{j(2\pi/N)(k-r)n} \right]$$

per cui, usando la (3.3), risulta

$$\sum_{n=0}^{N-1} \tilde{x}(n) e^{-j(2\pi/N)nr} = \tilde{X}(r)$$

Perciò i coefficienti $\tilde{X}(k)$ nella (3.2) sono dati da

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) e^{-j(2\pi/N)nk} \quad (3.4)$$

Notiamo che la sequenza $\tilde{X}(k)$ rappresentata dalla relazione (3.4) è periodica con periodo N , cioè $\tilde{X}(0) = \tilde{X}(N)$, $\tilde{X}(1) = \tilde{X}(N+1)$ ecc. Naturalmente, ciò è in accordo col fatto che gli esponenziali complessi rappresentati nell'espressione (3.1) sono distinti soltanto per $k = 0, 1, \dots, N-1$, e perciò nella rappresentazione di una sequenza periodica in serie di Fourier possono esservi solo N coefficienti distinti.

I coefficienti della serie di Fourier possono essere considerati come una sequenza di lunghezza finita, data dall'espressione (3.4) per $k = 0, \dots, N-1$ e zero per valori diversi di k , o come una sequenza periodica definita per ogni k dalla relazione (3.4). Chiaramente, queste due interpretazioni sono equivalenti. In generale è più conveniente interpretare i coefficienti della serie di Fourier $\tilde{X}(k)$ come una sequenza periodica. In questo modo si stabilisce una dualità tra i domini del tempo e della frequenza per la rappresentazione di sequenze periodiche in serie di Fourier. Le relazioni (3.2) e (3.4) possono essere considerate una coppia di trasformate e costituiscono la rappresentazione di una sequenza periodica in serie di Fourier discreta (DFS). Per comodità di rappresentazione queste espressioni saranno generalmente scritte in termini di W_N definito come

$$W_N = e^{-j(2\pi/N)}$$

Quindi le formule (di analisi e di sintesi) della DFS si scrivono come

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) W_N^{kn} \quad (3.5)$$

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) W_N^{-kn} \quad (3.6)$$

dove sia $\tilde{X}(k)$ che $\tilde{x}(n)$ sono sequenze periodiche.

La sequenza periodica $\tilde{X}(k)$ può essere utilmente interpretata come sequenza di campioni sul circolo unitario, equispaziati in angolo, della trasformata z di un periodo di $\tilde{x}(n)$. Per ottenere questa relazione, supponiamo

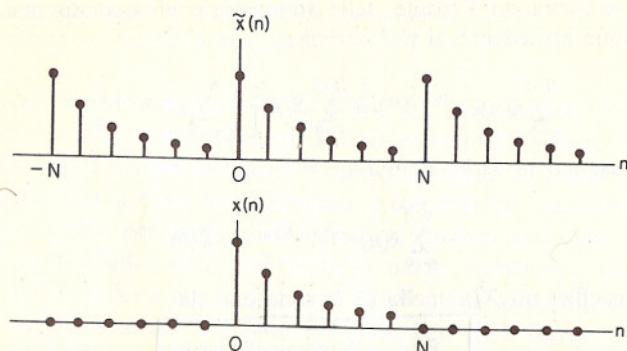


Fig. 3.1 Sequenza di lunghezza finita $x(n)$ coincidente con la sequenza periodica $\tilde{x}(n)$ per un periodo e zero altrove.

che $x(n)$ rappresenti un periodo di $\tilde{x}(n)$, cioè, $x(n) = \tilde{x}(n)$ per $0 \leq n \leq N-1$ e $x(n) = 0$ per valori diversi di n , come si può vedere in fig. 3.1. Quindi $X(z)$, la trasformata z di $x(n)$, è data da

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

oppure, poiché $x(n) = 0$ al di fuori dell'intervallo $0 \leq n \leq N-1$, da

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} \quad (3.7)$$

Confrontando le espressioni (3.5) e (3.7), si vede che tra $X(z)$ e $\tilde{X}(k)$ sussiste la relazione

$$\tilde{X}(k) = X(z) \Big|_{z=e^{j(2\pi/N)k}} = W_N^{-k} \quad (3.8)$$

Ciò corrisponde quindi a campionare la trasformata z , $X(z)$, in N punti egualmente spaziat in angolo sul circolo unitario, col primo campione situato in $z = 1$. Tali punti sul circolo unitario sono indicati in fig. 3.2.

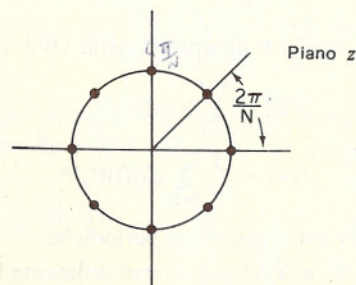


Fig. 3.2 Punti nel piano z in cui la trasformata z di un periodo di una sequenza periodica è uguale ai coefficienti della serie di Fourier.

ESEMPIO. Per illustrare la rappresentazione in serie di Fourier di una sequenza periodica, consideriamo la sequenza $\tilde{x}(n)$ di Fig. 3.3. Dalla (3.5) si deduce

$$\begin{aligned} \tilde{X}(k) &= \sum_{n=0}^4 W_{10}^{nk} = \sum_{n=0}^4 e^{-j(2\pi/10)nk} \\ &= e^{-j(4\pi k/10)} \frac{\sin(\pi k/2)}{\sin(\pi k/10)} \end{aligned} \quad (3.9)$$

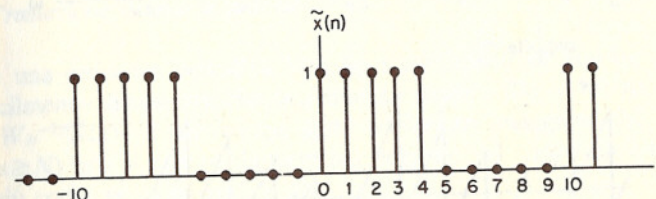


Fig. 3.3 Sequenza periodica per cui si vuole calcolare la rappresentazione in serie di Fourier.

Il modulo e la fase della sequenza periodica $\tilde{X}(k)$ data dalla (3.9) sono rappresentati in fig. 3.4. La trasformata z di un periodo di $\tilde{x}(n)$ valutata sul circolo unitario è

$$X(e^{j\omega}) = e^{-j2\omega} \frac{\sin(5\omega/2)}{\sin(\omega/2)}$$

Si verifica facilmente che la (3.8) è, in questo caso, soddisfatta. Il modulo e la fase di $X(e^{j\omega})$ sono rappresentati in fig. 3.5. È importante notare in particolare il fatto che le sequenze in fig. 3.4 (a) e (b) corrispondono, rispettivamente, a campioni di fig. 3.5 (a) e (b).

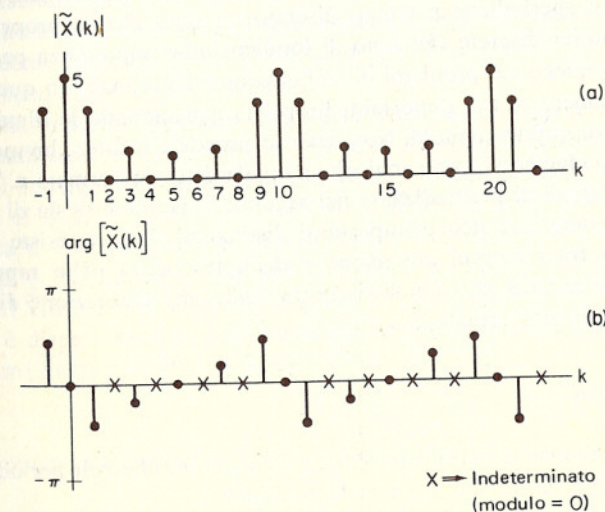


Fig. 3.4 Modulo e fase dei coefficienti della serie di Fourier della sequenza di fig. 3.3.

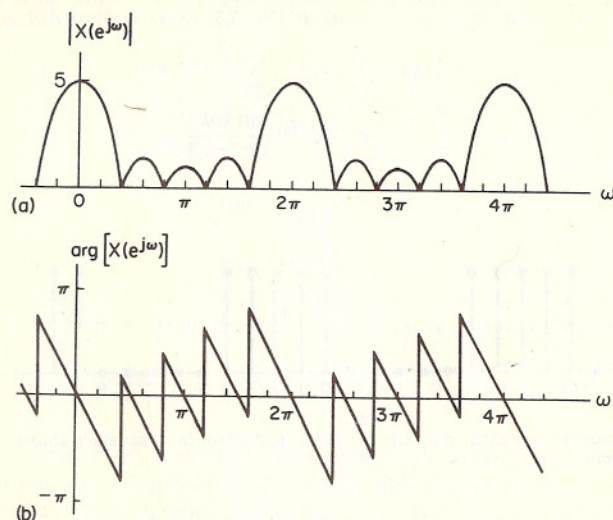


Fig. 3.5 Modulo e fase della trasformata z , valutata sul circolo unitario, di un periodo della sequenza di fig. 3.3.

3.2 PROPRIETÀ DELLA SERIE DI FOURIER DISCRETA

Analogamente a quanto si verifica con le trasformate di Fourier e di Laplace nel caso di segnali a tempo continuo, e con la trasformata z nel caso di sequenze aperiodiche a tempo discreto, ci sono alcune proprietà delle serie di Fourier discrete che sono di fondamentale importanza per un loro fruttuoso impiego nei problemi di elaborazione dei segnali. In questo paragrafo riassumeremo tali importanti proprietà, rimandando le dimostrazioni al probl. 2 di questo capitolo. Non deve sorprendere il fatto che molte delle proprietà fondamentali sono analoghe a quelle della trasformata z . Tuttavia, come si cercherà di puntualizzare nel seguito, la periodicità sia di $\tilde{x}(n)$ che di $\tilde{X}(k)$ dà luogo ad alcune importanti distinzioni. Inoltre, esiste una precisa dualità tra i domini del tempo e della frequenza nella rappresentazione mediante DFS, che non si riscontra nella rappresentazione di sequenze mediante la trasformata z .

3.2.1 Linearità

Se due sequenze periodiche $\tilde{x}_1(n)$ e $\tilde{x}_2(n)$, entrambe con periodi eguali ad N , vengono combinate secondo la relazione

$$\tilde{x}_3(n) = a\tilde{x}_1(n) + b\tilde{x}_2(n)$$

allora i coefficienti nella rappresentazione DFS di $\tilde{x}_3(n)$ sono dati da

$$\tilde{X}_3(k) = a\tilde{X}_1(k) + b\tilde{X}_2(k) \quad (3.10)$$

dove tutte le sequenze sono periodiche con periodo N .

3.2.2 Traslazione di una sequenza

Se una sequenza periodica $\tilde{x}(n)$ ha i coefficienti di Fourier $\tilde{X}(k)$, si può facilmente dimostrare che la sequenza traslata $\tilde{x}(n - m)$ ha i coefficienti $W_N^{-km}\tilde{X}(k)$. È ovvio che ogni traslazione maggiore del periodo (cioè $m \geq N$) non si può distinguere nel dominio del tempo da una traslazione più corta, $m' = m$ modulo N .

Poiché i coefficienti della serie di Fourier di una sequenza periodica costituiscono essi stessi una sequenza periodica, un risultato analogo si ottiene anche per una traslazione nei coefficienti di Fourier. In particolare, i valori della sequenza periodica $\tilde{X}(k + l)$ sono i coefficienti di Fourier della sequenza $W_N^{nl}\tilde{x}(n)$, dove l è un intero.

3.2.3 Proprietà di simmetria

Analogamente a quanto si è visto per la trasformata di Fourier nel cap. 1, esistono un certo numero di proprietà di simmetria anche per la rappresentazione mediante DFS di una sequenza periodica. Le dimostrazioni di queste proprietà sono di tipo simile a quelle del cap. 1 e vengono proposte come esercizio (v. il probl. 2 di questo capitolo). Le proprietà risultanti sono espone qui di seguito.

Se una generica sequenza complessa $\tilde{x}(n)$ ha coefficienti di Fourier $\tilde{X}(k)$, i coefficienti di Fourier per $\tilde{x}^*(n)$ sono $\tilde{X}^*(-k)$ e per $\tilde{x}^*(-n)$ sono $\tilde{X}^*(k)$. Di conseguenza, la DFS di $\text{Re}[\tilde{x}(n)]$ è $\tilde{X}_e(k)$, la parte simmetrica coniugata di $\tilde{X}(k)$, e la DFS di $j\text{Im}[\tilde{x}(n)]$ è $\tilde{X}_o(k)$, la parte coniugata antisimmetrica di $\tilde{X}(k)$. Inoltre, la DFS di $\tilde{x}_e(n)$ è $\text{Re}[\tilde{X}(k)]$ e la DFS di $\tilde{x}_o(n)$ è $j\text{Im}[\tilde{X}(k)]$. Ne consegue che per $\tilde{x}(n)$ reale $\text{Re}[\tilde{X}(k)]$ è una sequenza pari e $\text{Im}[\tilde{X}(k)]$ è una sequenza dispari. Anche il modulo di $\tilde{X}(k)$ è pari e la fase è dispari. Inoltre, per una sequenza reale, $\text{Re}[\tilde{X}(k)]$ è la DFS di $\tilde{x}_e(n)$ e $j\text{Im}[\tilde{X}(k)]$ è la DFS di $\tilde{x}_o(n)$.

3.2.4 Convoluzione periodica

Siano $\tilde{x}_1(n)$ e $\tilde{x}_2(n)$ due sequenze periodiche di periodo N con serie discrete di Fourier $\tilde{X}_1(k)$ e $\tilde{X}_2(k)$ rispettivamente. Si vuole determinare la

sequenza $\tilde{x}_3(n)$ la cui DFS è $\tilde{X}_1(k) \cdot \tilde{X}_2(k)$. Per dedurre questa relazione notiamo che è

$$\tilde{X}_1(k) = \sum_{m=0}^{N-1} \tilde{x}_1(m) W_N^{mk}$$

$$\tilde{X}_2(k) = \sum_{r=0}^{N-1} \tilde{x}_2(r) W_N^{rk}$$

per cui:

$$\tilde{X}_1(k) \tilde{X}_2(k) = \sum_{m=0}^{N-1} \sum_{r=0}^{N-1} \tilde{x}_1(m) \tilde{x}_2(r) W_N^{k(m+r)}$$

Risulta allora

$$\begin{aligned} \tilde{x}_3(n) &= \frac{1}{N} \sum_{k=0}^{N-1} W_N^{-nk} \tilde{X}_1(k) \tilde{X}_2(k) \\ &= \sum_{m=0}^{N-1} \tilde{x}_1(m) \sum_{r=0}^{N-1} \tilde{x}_2(r) \left[\frac{1}{N} \sum_{k=0}^{N-1} W_N^{-k(n-m-r)} \right] \end{aligned}$$

Consideriamo $\tilde{x}_3(n)$ per $0 \leq n \leq N-1$. Osserviamo che

$$\frac{1}{N} \sum_{k=0}^{N-1} W_N^{-k(n-m-r)} = \begin{cases} 1, & \text{per } r = (n-m) + lN \\ 0, & \text{altrove} \end{cases}$$

dove l è un intero qualunque. Ne risulta che

$$\tilde{x}_3(n) = \sum_{m=0}^{N-1} \tilde{x}_1(m) \tilde{x}_2(n-m) \quad (3.11a)$$

La (3.11 a) stabilisce che $\tilde{x}_3(n)$ si ottiene combinando $\tilde{x}_1(n)$ e $\tilde{x}_2(n)$ in un modo che ricorda una convoluzione. È importante notare, tuttavia, che, contrariamente alla convoluzione di sequenze aperiodiche, le sequenze $\tilde{x}_1(m)$ e $\tilde{x}_2(n-m)$ della relazione (3.11 a) sono periodiche in m con periodo N e di conseguenza lo è anche il loro prodotto. Inoltre, la somma è eseguita solo su un periodo. Questo tipo di convoluzione è comunemente chiamata convoluzione periodica. Cambiando semplicemente l'indice della somma nella relazione (3.11 a) si può mostrare che risulta

$$\tilde{x}_3(n) = \sum_{m=0}^{N-1} \tilde{x}_2(m) \tilde{x}_1(n-m) \quad (3.11b)$$

La fig. 3.6 illustra il procedimento della formazione della convoluzione periodica di due sequenze periodiche corrispondente alla relazione (3.11 a). In questo tipo di convoluzione, accade che quando un periodo esce dall'intervallo su cui essa è valutata, vi entra il periodo successivo.

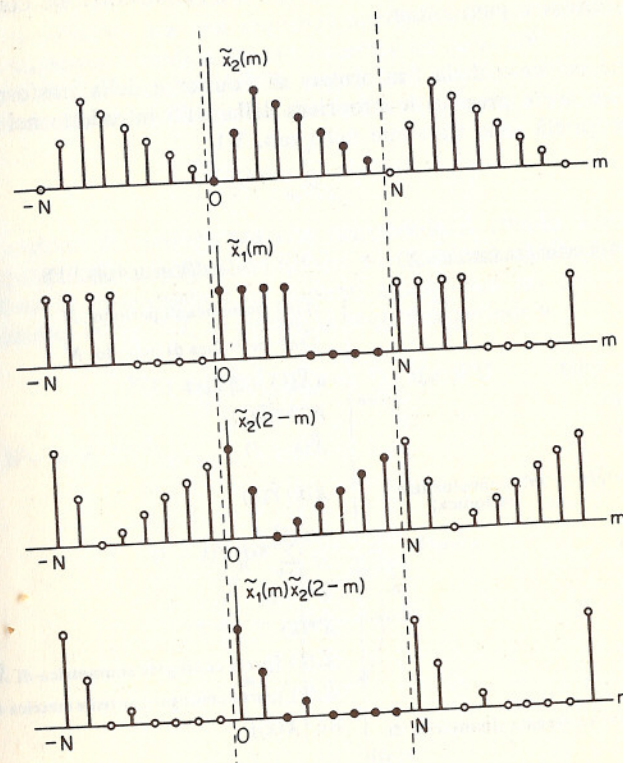


Fig. 3.6 Procedimento di formazione della convoluzione periodica di due sequenze periodiche.

Scambiando i ruoli giocati dalle variabili tempo e frequenza, otteniamo un risultato quasi identico al precedente. Cioè, risulta che la sequenza periodica

$$\tilde{x}_3(n) = \tilde{x}_1(n) \tilde{x}_2(n)$$

dove $\tilde{x}_1(n)$ e $\tilde{x}_2(n)$ sono sequenze periodiche di periodo N , ha i coefficienti di Fourier dati da

$$\tilde{X}_3(k) = \frac{1}{N} \sum_{l=0}^{N-1} \tilde{X}_1(l) \tilde{X}_2(k-l) \quad (3.12)$$

corrispondenti alla convoluzione periodica di $\tilde{X}_1(k)$ e $\tilde{X}_2(k)$ moltiplicata per $1/N$.

3.3 RIASSUNTO DELLE PROPRIETÀ DELLA RAPPRESENTAZIONE CON LA DFS DI SEQUENZE PERIODICHE

Come nel caso della trasformata di Fourier e della trasformata z , è spesso utile avere presenti le proprietà della DFS introdotte nel par. 3.2. Queste proprietà sono riassunte nella tab. 3.1.

Tab. 3.1

Sequenza periodica (periodo N)	Coefficienti della DFS
1. $\tilde{x}(n)$	$\tilde{X}(k)$ periodica di periodo N
2. $\tilde{y}(n)$	$\tilde{Y}(k)$ periodica di periodo N
3. $a\tilde{x}(n) + b\tilde{y}(n)$	$a\tilde{X}(k) + b\tilde{Y}(k)$
4. $\tilde{x}(n+m)$	$W_N^{-km}\tilde{X}(k)$
5. $W_N^{ln}\tilde{x}(n)$	$\tilde{X}(k+l)$
6. $\sum_{m=0}^{N-1} \tilde{x}(m)\tilde{y}(n-m)$ (convoluzione periodica)	$\tilde{X}(k)\tilde{Y}(k)$
7. $\tilde{x}(n)\tilde{y}(n)$	$\frac{1}{N} \sum_{l=0}^{N-1} \tilde{X}(l)\tilde{Y}(k-l)$
8. $\tilde{x}^*(n)$	$\tilde{X}^*(-k)$
9. $\tilde{x}^*(-n)$	$\tilde{X}^*(k)$
10. $\text{Re} [\tilde{x}(n)]$	$\tilde{X}_e(k)$ [parte coniugata simmetrica di $\tilde{X}(k)$]
11. $j \text{Im} [\tilde{x}(n)]$	$\tilde{X}_o(k)$ [parte coniugata antisimmetrica di $\tilde{X}(k)$]
12. $\tilde{x}_e(n)$ [parte coniugata simmetrica di $\tilde{x}(n)$]	$\text{Re} [\tilde{X}(k)]$
13. $\tilde{x}_o(n)$ [parte coniugata antisimmetrica di $\tilde{x}(n)$]	$j \text{Im} [\tilde{X}(k)]$

Le proprietà seguenti valgono solo quando $\tilde{x}(n)$ è reale:

14. $\tilde{x}(n)$ reale qualsiasi	$\begin{cases} \tilde{X}(k) = \tilde{X}^*(-k) \\ \text{Re} [\tilde{X}(k)] = \text{Re} [\tilde{X}(-k)] \\ \text{Im} [\tilde{X}(k)] = -\text{Im} [\tilde{X}(-k)] \\ \tilde{X}(k) = \tilde{X}(-k) \\ \arg [\tilde{X}(k)] = -\arg [\tilde{X}(-k)] \end{cases}$
15. $\tilde{x}_e(n)$	$\text{Re} [\tilde{X}(k)]$
16. $\tilde{x}_o(n)$	$j \text{Im} [\tilde{X}(k)]$

3.4 CAMPIONAMENTO DELLA TRASFORMATA z

Abbiamo visto nel par. 3.1 che i valori $\tilde{X}(k)$ nella rappresentazione mediante la DFS di una sequenza periodica coincidono con i valori della trasformata z di un singolo periodo di $\tilde{x}(n)$ presi in N punti spazati uni-

formemente sul circolo unitario. In questo paragrafo esaminiamo, più in generale, la relazione tra una sequenza aperiodica la cui trasformata z è $X(z)$ e la sequenza periodica per cui i coefficienti della DFS corrispondono a campioni di $X(z)$ equispaziati in angolo sul circolo unitario. A questo scopo, consideriamo una sequenza aperiodica $x(n)$ la cui trasformata z

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (3.13)$$

ha una regione di convergenza che comprende il circolo unitario, condizione questa che è sempre soddisfatta per sequenze di lunghezza finita. Se si valuta la trasformata z in N punti equispaziati sul circolo unitario come mostrato in fig. 3.7, si ottiene la sequenza periodica

$$\tilde{X}(k) = X(z)|_{z=W_N^{-k}} = \sum_{n=-\infty}^{\infty} x(n)W_N^{kn} \quad (3.14)$$

dove è $W_N = e^{-j(2\pi/N)}$.

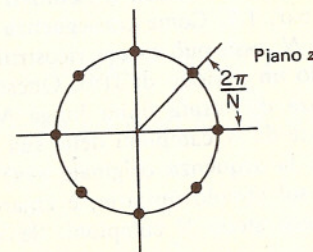


Fig. 3.7 Punti del circolo unitario dove è campionata $X(z)$ per ottenere una sequenza periodica $\tilde{X}(k)$.

Abbiamo appena visto che esiste una relazione univoca tra una sequenza periodica $\tilde{X}(k)$ e la sequenza periodica $\tilde{x}(n)$ ottenuta come

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k)W_N^{-kn} \quad (3.15)$$

Per ricavare la relazione tra la sequenza periodica $\tilde{x}(n)$ e la sequenza originaria $x(n)$, sostituiamo i valori di $\tilde{X}(k)$ dall'espressione (3.14) nella (3.15), ottenendo

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{m=-\infty}^{\infty} x(m)W_N^{km}W_N^{-kn}$$

Scambiando l'ordine delle sommatorie si ha

$$\tilde{x}(n) = \sum_{m=-\infty}^{\infty} x(m) \left[\frac{1}{N} \sum_{k=0}^{N-1} W_N^{-k(n-m)} \right]$$

Usando la proprietà (3.3) si ricava

$$\frac{1}{N} \sum_{k=0}^{N-1} W_N^{-k(n-m)} = 1 \quad \text{per} \quad m = n + rN$$

e zero altrove, per cui risulta

$$\tilde{x}(n) = \sum_{r=-\infty}^{\infty} x(n + rN) \quad (3.16)$$

Perciò, la sequenza periodica risultante è originata dalla sequenza aperiodica sovrapponendo ripetizioni successive di quest'ultima. Questo ricorda la relazione ricavata nel par. 1.7 tra la trasformata di Fourier di un segnale a tempo continuo e la trasformata di Fourier del segnale a tempo discreto ottenuto con il campionamento periodico: basta notare la somiglianza delle espressioni (1.29) e (3.16). Dalla relazione (3.16) si vede che se la sequenza aperiodica $x(n)$ è di durata finita inferiore a N , allora ogni periodo di $\tilde{x}(n)$ è una replica di $x(n)$; se invece è di durata maggiore di N , c'è una sovrapposizione di valori diversi da zero, con il risultato di un *aliasing* simile a quello considerato nel par. 1.7. Come conseguenza si ha che se $x(n)$ è di durata finita minore di N , essa può essere ricostruita esattamente da $\tilde{x}(n)$ semplicemente estraendo un periodo di $\tilde{x}(n)$. Questo è equivalente a dire, allora, che una sequenza di durata finita lunga N (o meno) può essere rappresentata esattamente da N campioni della sua trasformata z presi sul cerchio unitario. Poiché la sequenza originale $x(n)$ può essere ricostruita dagli N valori di $X(z)$ sul cerchio unitario, è chiaro che anche $X(z)$ può essere riottenuta da questi stessi N campioni. Se $x(n)$ è zero per $n \geq N$, allora si ha

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \quad (3.17)$$

Poiché $x(n) = \tilde{x}(n)$ per $0 \leq n \leq N-1$, si può sostituire la relazione (3.15) nella (3.17), ottenendo

$$X(z) = \sum_{n=0}^{N-1} \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) W_N^{-kn} z^{-n}$$

Scambiando l'ordine delle sommatorie si ha

$$X(z) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) \left[\sum_{n=0}^{N-1} (W_N^{-k} z^{-1})^n \right]$$

che può essere scritta come

$$\begin{aligned} X(z) &= \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) \frac{1 - z^{-N}}{1 - W_N^{-k} z^{-1}} \\ &= \frac{1 - z^{-N}}{N} \sum_{k=0}^{N-1} \frac{\tilde{X}(k)}{1 - W_N^{-k} z^{-1}} \end{aligned} \quad (3.18)$$

Questa relazione esprime $X(z)$, la trasformata z di una sequenza di durata finita lunga N , in termini di N « campioni frequenziali » di $X(z)$ sul cerchio unitario. Come si vedrà in un capitolo successivo, questa espressione è la base per una possibile realizzazione di un sistema avente risposta all'impulso di durata finita. Con la sostituzione $z = e^{j\omega}$ si può dimostrare che la formula (3.18) diventa

$$X(e^{j\omega}) = \sum_{k=0}^{N-1} \tilde{X}(k) \Phi\left(\omega - \frac{2\pi}{N} k\right) \quad (3.19)$$

dove

$$\Phi(\omega) = \frac{\sin(\omega N/2)}{N \sin(\omega/2)} e^{-j\omega[(N-1)/2]} \quad (3.20)$$

La funzione $\sin(\omega N/2)/[N \sin(\omega/2)]$ è riportata nella fig. 3.8 per $N = 5$. Si noti che la funzione $\Phi(\omega)$ ha la proprietà

$$\Phi\left(\frac{2\pi}{N} k\right) = \begin{cases} 0, & k = 1, 2, \dots, N-1 \\ 1, & k = 0 \end{cases}$$

per cui risulta

$$X(e^{j\omega})|_{\omega=(2\pi/N)k} = \tilde{X}(k), \quad k = 0, 1, \dots, N-1 \quad (3.21)$$

cioè l'interpolazione è esatta nei punti di campionamento originali, come è lecito attendersi.

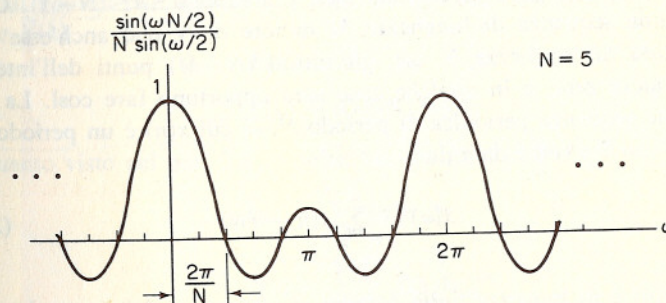


Fig. 3.8 Grafico della funzione $\sin(\omega N/2)/[N \sin(\omega/2)]$, definita nella relazione (3.20), per $N = 5$.

3.5 RAPPRESENTAZIONE DI FOURIER PER SEQUENZE DI DURATA FINITA - LA TRASFORMATA DI FOURIER DISCRETA

Nel paragrafo precedente abbiamo considerato la rappresentazione di sequenze periodiche in termini della serie di Fourier discreta. La stessa rappresentazione può essere applicata a sequenze di durata finita, purché la si interpreti correttamente. La rappresentazione di Fourier che ne risulta verrà indicata come la trasformata di Fourier discreta (DFT).

I risultati del paragrafo precedente suggeriscono due punti di vista per la rappresentazione di Fourier di sequenze con durata finita. In particolare, possiamo rappresentare una sequenza di durata finita lunga N con una sequenza periodica di periodo N , il cui andamento nel periodo sia identico alla sequenza di durata finita². Come la sequenza periodica ha una rappresentazione unica con la DFS, la stessa cosa vale per la sequenza di durata finita originaria, in quanto dalla DFS si può calcolare un singolo periodo della sequenza periodica, e quindi la sequenza di durata finita.

Un punto di vista alternativo è suggerito nel paragrafo precedente, là dove si è mostrato che una sequenza di durata finita può essere rappresentata esattamente dai campioni della sua trasformata z . In particolare abbiamo dimostrato che la sequenza periodica ottenuta campionando la trasformata z in N punti equispaziati sul circolo unitario è identica ai coefficienti della serie di Fourier discreta della sequenza periodica costruita come prima esposto. Si è visto che la sequenza corrispondente a questi campioni della trasformata z è una versione della sequenza originale ripetuta periodicamente, in modo che non vi è *aliasing* quando si usano N campioni della trasformata z . Perciò entrambi i punti di vista conducono alla rappresentazione di una sequenza di durata finita come un periodo di una sequenza periodica.

Consideriamo una sequenza di durata finita $x(n)$ di lunghezza N in modo che $x(n) = 0$ eccetto che nell'intervallo $0 \leq n \leq (N-1)$. Chiaramente, una sequenza di lunghezza M minore di N può anch'essa essere considerata di lunghezza N con gli ultimi $(N-M)$ punti dell'intervallo aventi valore zero, e in qualche caso sarà opportuno fare così. La corrispondente sequenza periodica di periodo N , di cui $x(n)$ è un periodo, sarà indicata con $\tilde{x}(n)$ ed è data da

$$\tilde{x}(n) = \sum_{r=-\infty}^{\infty} x(n + rN) \quad (3.22a)$$

Poiché $x(n)$ è di lunghezza finita N , non vi è sovrapposizione tra i termini $x(n + rN)$ per valori di r differenti. Perciò la relazione (3.22 a) può essere scritta nella forma alternativa³

$$\tilde{x}(n) = x(n \text{ modulo } N) \quad (3.22b)$$

² Per semplicità si assume in generale che il campo in cui la sequenza è diversa da zero sia $0 \leq n \leq N-1$; questo è però arbitrario e si possono derivare risultati validi per qualsiasi intervallo di N campioni.

³ Se n è espresso come $n = n_1 + n_2N$ con $0 \leq n_1 \leq N-1$, n modulo N è uguale a n_1 .

Per comodità useremo la notazione $((n))_N$ per indicare « n modulo N » e con questa notazione l'espressione (3.22 b) diventa

$$\tilde{x}(n) = x((n))_N \quad (3.23a)$$

La sequenza di durata finita $x(n)$ si ricava da $\tilde{x}(n)$ estraendone un periodo, cioè

$$x(n) = \begin{cases} \tilde{x}(n), & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

Ancora per comodità di notazione è utile introdurre la sequenza rettangolare $\mathcal{R}_N(n)$ definita come

$$\mathcal{R}_N(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

Con questa simbologia la relazione precedente può essere espressa come

$$x(n) = \tilde{x}(n)\mathcal{R}_N(n) \quad (3.23b)$$

In base alla definizione del par. 3.1, i coefficienti della serie di Fourier discreta $\tilde{X}(k)$ relativi alla sequenza periodica $\tilde{x}(n)$ sono anch'essi una sequenza periodica di periodo N . Per mantenere la dualità tra i domini del tempo e della frequenza, sceglieremo i coefficienti di Fourier da associare a una sequenza di durata finita come la sequenza di durata finita costituita da un periodo di $\tilde{X}(k)$. Perciò, se indichiamo con $X(k)$ i coefficienti di Fourier che associamo a $x(n)$, vale la seguente relazione che lega $X(k)$ e $\tilde{X}(k)$

$$\tilde{X}(k) = X((k))_N \quad (3.24a)$$

$$X(k) = \tilde{X}(k)\mathcal{R}_N(k) \quad (3.24b)$$

Da quanto visto nel par. 3.1, la relazione tra $\tilde{X}(k)$ e $\tilde{x}(n)$ è

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n)W_N^{kn} \quad (3.25a)$$

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k)W_N^{-kn} \quad (3.25b)$$

Poiché le somme nelle (3.25 a) e (3.25 b) riguardano solo l'intervallo tra 0 e $(N-1)$, segue dalle (3.23), (3.24) e (3.25) che

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n)W_N^{kn}, & 0 \leq k \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (3.26a)$$

$$x(n) = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} X(k)W_N^{-kn}, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (3.26b)$$

La coppia di trasformate indicate nelle (3.26) sarà chiamata trasformata di Fourier discreta (DFT): la relazione (3.26 a) rappresenta l'analisi e la (3.26 b) la sintesi della sequenza $x(n)$. Si può notare che sulla base dei ragionamenti del par. 3.4, la DFT di una sequenza di durata finita corrisponde a campioni della sua trasformata z equispaziati sul circolo unitario. Va sottolineato che, proprio come è di scarsa rilevanza la distinzione tra sequenza di durata finita lunga N e sequenza periodica di periodo N , nel senso che entrambe sono definite da N valori, così anche la distinzione tra le espressioni (3.25) e (3.26) è relativamente poco importante. Tuttavia, come vedremo nel prossimo paragrafo, è sempre importante ricordare che quando si parla di relazioni tramite la DFT, una sequenza di lunghezza finita è rappresentata come un periodo di una sequenza periodica.

3.6 PROPRIETÀ DELLA TRASFORMATTA DI FOURIER DISCRETA

In questo paragrafo prendiamo in esame alcune proprietà della DFT per sequenze di durata finita. Queste proprietà sono sostanzialmente analoghe a quelle presentate nel par. 3.2 per sequenze periodiche e derivano dalla periodicità *implicita* nella rappresentazione di sequenze di durata finita con la DFT. Il nostro scopo in questo paragrafo è di riesaminare queste proprietà con riferimento ad una sequenza di lunghezza finita definita solo nell'intervallo $0 \leq n \leq N-1$.

3.6.1 Linearità

Se due sequenze di durata finita $x_1(n)$ e $x_2(n)$ sono combinate linearmente, come

$$x_3(n) = ax_1(n) + bx_2(n)$$

allora la DFT di $x_3(n)$ è

$$X_3(k) = aX_1(k) + bX_2(k)$$

È chiaro che se $x_1(n)$ ha durata N_1 e $x_2(n)$ ha durata N_2 , allora la durata massima di $x_3(n)$ sarà $N_3 = \max[N_1, N_2]$. Perciò, in generale, le DFT devono essere calcolate con $N = N_3$. Se, ad esempio, $N_1 < N_2$, allora $X_1(k)$ è la DFT della sequenza $x_1(n)$ allungata con $N_2 - N_1$ zeri. In altri termini, sarà

$$X_1(k) = \sum_{n=0}^{N_1-1} x_1(n) W_{N_2}^{kn}, \quad 0 \leq k \leq N_2 - 1$$

$$X_2(k) = \sum_{n=0}^{N_2-1} x_2(n) W_{N_2}^{kn}, \quad 0 \leq k \leq N_2 - 1$$

3.6.2 Traslazione circolare di una sequenza

Consideriamo una sequenza $x(n)$ come quella di fig. 3.9 (a), la sua versione periodica $\tilde{x}(n)$ come mostrato in fig. 3.9 (b), e $\tilde{x}(n+m)$, cioè la sequenza ottenuta traslando $\tilde{x}(n)$ di m campioni, come indicato nella fig. 3.9 (c). La sequenza di durata finita, che chiameremo $x_1(n)$, ottenuta estraendo un periodo di $\tilde{x}(n+m)$ nell'intervallo $0 \leq n \leq N-1$, è mostrata in fig. 3.9 (d). Il confronto della fig. 3.9 (a) con la 3.9 (d) indica chiaramente che $x_1(n)$ non corrisponde a una traslazione lineare di $x(n)$, tanto è vero che entrambe le sequenze sono limitate all'intervallo tra 0 e $(N-1)$. Con riferimento alle fig. 3.9 (b) e (c), si vede che traslando la sequenza periodica e prendendo in esame l'intervallo tra 0 e $(N-1)$, quando un campione esce da questo intervallo, un campione identico vi rientra dall'altra parte. Allora si può immaginare di costruire $x_1(n)$ traslando $x(n)$ in modo tale che ogni campione che esce dall'intervallo tra 0 e $(N-1)$ da una parte vi rientri dall'altra.

Un'interpretazione utile di questa traslazione consiste nell'immaginare la sequenza di durata finita $x(n)$ disposta lungo la circonferenza di un cilindro in modo che il cilindro abbia una circonferenza di esattamente N punti. Se si percorre più volte la circonferenza del cilindro, la sequenza che si vede è proprio la sequenza periodica $\tilde{x}(n)$. Una traslazione lineare della sequenza periodica $\tilde{x}(n)$ corrisponde allora a una *rotazione* del cilindro. Questo tipo di traslazione di una sequenza è generalmente chiamato traslazione circolare. Per esprimere in maniera più formale la traslazione circolare possiamo usare le relazioni (3.23 a) e (3.23 b). Nella fattispecie risulta

$$\tilde{x}_1(n) = \tilde{x}(n+m) = x((n+m))_N$$

Perciò, dall'espressione (3.23 b) si ha

$$x_1(n) = x((n+m))_N \mathcal{R}_N(n)$$

Vogliamo adesso porre in relazione la DFT di $x(n)$ e la DFT di $x_1(n)$. Con riferimento al par. 3.2.2 ricordiamo che se $\tilde{X}(k)$ e $\tilde{X}_1(k)$ indicano le DFS delle sequenze periodiche $\tilde{x}(n)$ e $\tilde{x}_1(n) = \tilde{x}(n+m)$, vale la

$$\tilde{X}_1(k) = W_N^{-km} \tilde{X}(k) \quad (3.27)$$

Di conseguenza, dalla (3.24 b) segue che

$$X_1(k) = W_N^{-km} X(k) \quad (3.28)$$

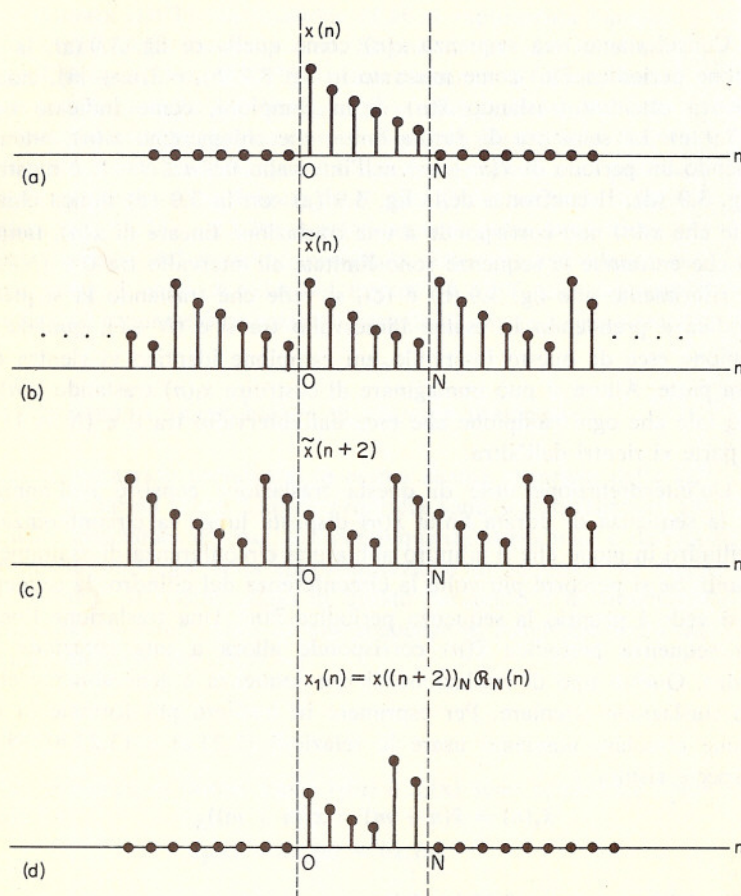


Fig. 3.9 Traslazione circolare di una sequenza.

A causa della dualità tra il dominio del tempo e quello della frequenza, un risultato analogo vale quando si applica una traslazione circolare ai coefficienti della DFT. In particolare, indicando con $X(k)$ e $X_1(k)$ rispettivamente la DFT di $x(n)$ e di $x_1(n)$, se è

$$X_1(k) = X((k+1))_N \mathcal{R}_N(k) \quad (3.29)$$

allora risulta

$$x_1(n) = W_N^{ln} x(n) \quad (3.30)$$

3.6.3 Proprietà di simmetria

Nel cap. 1 si è considerata la scomposizione di una generica sequenza nella somma delle sue componenti simmetrica coniugata e antisimmetrica coniugata ed è stato presentato un insieme di relazioni di simmetria per la trasformata di Fourier. Nello studio delle proprietà di simmetria della DFT per sequenze di durata finita, non si possono usare in generale le definizioni di componenti coniugata simmetrica e coniugata antisimmetrica nella forma data nel par. 1.6, perché, per una data sequenza $x(n)$ di durata N , la componente simmetrica coniugata $x_e(n)$ e la componente antisimmetrica coniugata $x_o(n)$ sono entrambe di lunghezza $(2N-1)$. Si può osservare, in ogni modo, che, per una sequenza periodica $\tilde{x}(n)$ di periodo N , le componenti simmetrica coniugata e antisimmetrica coniugata sono ancora periodiche con periodo N . Questo fatto suggerisce la scomposizione di $x(n)$ in due sequenze finite di durata N , corrispondenti ad un periodo delle componenti coniugate di $\tilde{x}(n)$. Indicheremo queste componenti di $x(n)$ con $x_{ep}(n)$ e $x_{op}(n)$. Perciò, ponendo

$$\tilde{x}(n) = x((n))_N \quad (3.31)$$

e

$$\tilde{x}_e(n) = \frac{1}{2}[\tilde{x}(n) + \tilde{x}^*(-n)] \quad (3.32)$$

e

$$\tilde{x}_o(n) = \frac{1}{2}[\tilde{x}(n) - \tilde{x}^*(-n)] \quad (3.33)$$

definiamo $x_{ep}(n)$ e $x_{op}(n)$ come

$$x_{ep}(n) = \tilde{x}_e(n) \mathcal{R}_N(n) \quad (3.34a)$$

$$x_{op}(n) = \tilde{x}_o(n) \mathcal{R}_N(n) \quad (3.34b)$$

oppure, in modo del tutto equivalente,

$$x_{ep}(n) = \frac{1}{2}[x((n))_N + x^*((-n))_N] \mathcal{R}_N(n) \quad (3.35a)$$

$$x_{op}(n) = \frac{1}{2}[x((n))_N - x^*((-n))_N] \mathcal{R}_N(n) \quad (3.35b)$$

Chiaramente $x_{ep}(n)$ e $x_{op}(n)$ non equivalgono alle sequenze $x_e(n)$ e $x_o(n)$ come sono state definite dalla (1.22). Tuttavia si può dimostrare (v. probl. 17 di questo capitolo) che valgono le relazioni

$$x_{ep}(n) = [x_e(n) + x_e(n-N)] \mathcal{R}_N(n) \quad (3.36a)$$

e

$$x_{op}(n) = [x_o(n) + x_o(n-N)] \mathcal{R}_N(n) \quad (3.36b)$$

In altre parole, $x_{ep}(n)$ e $x_{op}(n)$ possono essere viste come generate da un *aliasing* delle $x_e(n)$ e $x_o(n)$ nell'intervallo $0 \leq n \leq N-1$. Le sequenze $x_{ep}(n)$

e $x_{op}(n)$ sono dette *componenti simmetrica coniugata periodica e antisimmetrica coniugata periodica* di $x(n)$. Quando $x_{ep}(n)$ e $x_{op}(n)$ sono reali, vengono dette, rispettivamente, componente periodica pari e periodica dispari. La scelta di questa terminologia è in qualche modo fuorviante poiché le sequenze $x_{ep}(n)$ e $x_{op}(n)$ non sono sequenze periodiche, ma rappresentano un periodo delle sequenze periodiche $\tilde{x}_e(n)$ e $\tilde{x}_o(n)$.

Le definizioni (3.35 a) e (3.35 b) esprimono $x_{ep}(n)$ e $x_{op}(n)$ in funzione di $x(n)$; la relazione inversa, cioè l'espressione di $x(n)$ in funzione di $x_{ep}(n)$ e di $x_{op}(n)$, può essere ottenuta a partire dalle (3.32) e (3.33) da cui si ricava

$$\tilde{x}(n) = \tilde{x}_e(n) + \tilde{x}_o(n)$$

È quindi

$$x(n) = \tilde{x}(n)R_N(n) = [\tilde{x}_e(n) + \tilde{x}_o(n)]R_N(n) = \tilde{x}_e(n)R_N(n) + \tilde{x}_o(n)R_N(n) \quad (3.37)$$

Combinando le espressioni (3.34) e (3.37) si ottiene

$$x(n) = x_{ep}(n) + x_{op}(n) \quad (3.38)$$

Le proprietà di simmetria della DFT possono ora essere ricavate in maniera immediata applicando i risultati del par. 3.2.3. Consideriamo quindi una sequenza finita $x(n)$ di durata N la cui DFT sia $X(k)$. Allora la DFT di $x^*(n)$ sarà $X^*((-k))_N R_N(k)$ e quella di $x^*((-n))_N R_N(n)$ sarà $X^*(k)$. La DFT di $\text{Re}[x(n)]$ sarà $X_{ep}(k)$ e la DFT di $j\text{Im}[x(n)]$ $X_{op}(k)$; analogamente la DFT di $x_{ep}(n)$ è $\text{Re}[X(k)]$ e la DFT di $x_{op}(n)$ è $j\text{Im}[X(k)]$. Da ciò si ricava che, per $x(n)$ reale, $\text{Re}[X(k)]$ e $|X(k)|$ sono sequenze periodiche pari e $\text{Im}[X(k)]$ e $\arg[X(k)]$ sono sequenze periodiche dispari. Inoltre, per una sequenza reale, $\text{Re}[X(k)]$ è la DFT di $x_{ep}(n)$ e $j\text{Im}[X(k)]$ è la DFT di $x_{op}(n)$.

3.6.4 Convoluzione circolare

Nel par. 3.2.4 abbiamo osservato che la moltiplicazione dei coefficienti delle DFS di due sequenze corrisponde alla convoluzione periodica delle sequenze. Qui consideriamo le sequenze $x_1(n)$ e $x_2(n)$, entrambe finite con durata N , con le loro DFT $X_1(k)$ e $X_2(k)$, e vogliamo determinare la sequenza $x_3(n)$ i cui coefficienti della DFT siano $X_1(k)X_2(k)$. Per determinare $x_3(n)$ basta semplicemente applicare i risultati del par. 3.2.4; più

precisamente, $x_3(n)$ corrisponde ad un periodo di $\tilde{x}_3(n)$, che è data dalla (3.11). Pertanto si ha

$$\begin{aligned} x_3(n) &= \left[\sum_{m=0}^{N-1} \tilde{x}_1(m) \tilde{x}_2(n-m) \right] R_N(n) \\ &= \left[\sum_{m=0}^{N-1} x_1((m))_N x_2((n-m))_N \right] R_N(n) \end{aligned} \quad (3.39)$$

La (3.39) differisce in alcuni importanti aspetti dalla convoluzione lineare di $x_1(n)$ e $x_2(n)$ così come è stata definita dalla (1.7). Nel caso della *convoluzione lineare*, l'operazione fondamentale è data dalla moltiplicazione di $x_1(n)$ per una replica di $x_2(n)$ rovesciata e traslata linearmente e dalla successiva somma dei prodotti; per ottenere i diversi valori della sequenza che rappresenta la convoluzione, le due sequenze devono essere traslate successivamente l'una rispetto all'altra. Al contrario, per la convoluzione espressa dalla (3.39), si può immaginare di disporre una delle due sequenze lungo la circonferenza di un cilindro che abbia una circonferenza di esattamente N punti. La seconda sequenza viene rovesciata nel tempo e disposta anch'essa lungo la circonferenza di un cilindro con circonferenza di N punti. Se si immagina di porre un cilindro dentro l'altro, allora i successivi valori della convoluzione si ottengono moltiplicando i valori su un cilindro per i corrispondenti valori sull'altro e sommando poi gli N prodotti risultanti; per generare valori successivi della convoluzione occorre ruotare un cilindro rispetto all'altro. Una breve riflessione dovrebbe chiarire che questa convoluzione è del tutto equivalente a costruire dapprima le due sequenze periodiche e poi farne la convoluzione nel modo descritto dalla (3.11) ed illustrato in fig. 3.6. Con questo tipo di interpretazione, tale convoluzione viene spesso indicata col nome di *convoluzione circolare*. La convoluzione circolare su N punti di due sequenze $x_1(n)$ e $x_2(n)$ è spesso rappresentata con la notazione $x_1(n) \circledast x_2(n)$.

ESEMPIO. Un semplice esempio di convoluzione circolare è ricavabile dai risultati del par. 3.6.2. Sia $x_2(n)$ una sequenza di durata finita N , e valga

$$x_1(n) = \delta(n - n_0)$$

dove $n_0 < N$. Chiaramente $x_1(n)$ può essere considerata come la sequenza di durata finita definita da

$$x_1(n) = \begin{cases} 0, & 0 \leq n < n_0 \\ 1, & n = n_0 \\ 0, & n_0 < n \leq N-1 \end{cases}$$

La DFT di $x_1(n)$ è

$$X_1(k) = W_N^{kn_0}$$

Se formiamo il prodotto

$$X_3(k) = W_N^{kn_0} X_2(k)$$

si può osservare, facendo riferimento al par. 3.6.2, che la sequenza di durata finita corrispondente a $X_3(k)$ è la sequenza $x_2(n)$ ruotata di n_0 campioni a destra nell'intervallo $0 \leq n \leq N-1$. Vale a dire, la convoluzione circolare di una sequenza $x_2(n)$ con un campione unitario ritardato risulta nella rotazione della sequenza $x_2(n)$ nell'intervallo $0 \leq n \leq N-1$. Questo esempio è illustrato in fig. 3.10 per il caso di $N=5$ e $n_0=1$. Sono riportate le sequenze $x_2(m)$ e $x_1(m)$ seguite da $x_2((0-m))_N$ e $x_2((1-m))_N$; l'ultima figura mostra il risultato della convoluzione circolare di $x_1(n)$ e $x_2(n)$.

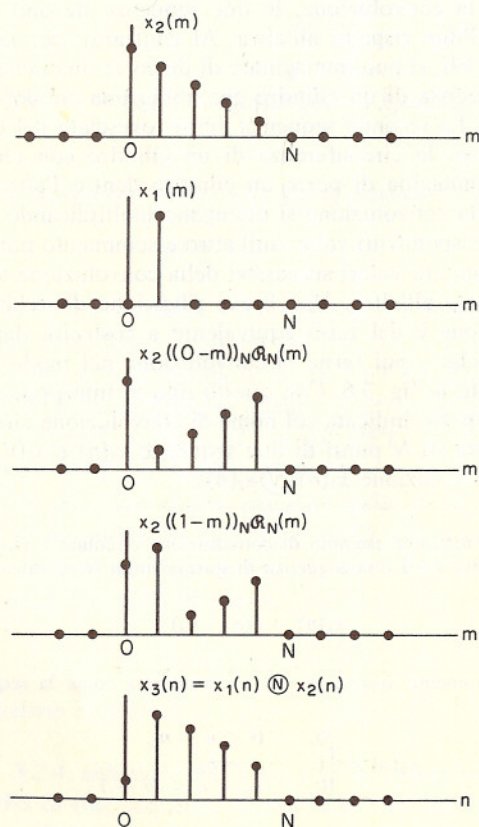


Fig. 3.10 Convoluzione circolare di due sequenze.

ESEMPIO. Come altro esempio di convoluzione circolare si consideri

$$x_1(n) = x_2(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

E allora

$$X_1(k) = X_2(k) = \sum_{n=0}^{N-1} W_N^{kn} = \begin{cases} N, & k=0 \\ 0, & \text{altrove} \end{cases}$$

Perciò risulta

$$X_3(k) = X_1(k)X_2(k) = \begin{cases} N^2, & k=0 \\ 0, & \text{altrove} \end{cases}$$

e si vede che

$$x_3(n) = N, \quad 0 \leq n \leq N-1$$

Tutto questo è schematizzato in fig. 3.11. Chiaramente, al ruotare della sequenza $x_2(n)$ rispetto a $x_1(n)$, la somma dei prodotti $x_1(m)x_2(n-m)$ è sempre uguale a N , come si può vedere in fig. 3.11. Naturalmente è possibile considerare $x_1(n)$ e $x_2(n)$ come sequenze di $2N$ punti aggiungendo N zeri; se ora si esegue la convoluzione circolare delle sequenze allungate, si ottiene la sequenza di fig. 3.12, che, come si può vedere, è identica alla convoluzione lineare delle sequenze di durata finita $x_1(n)$ e $x_2(n)$.

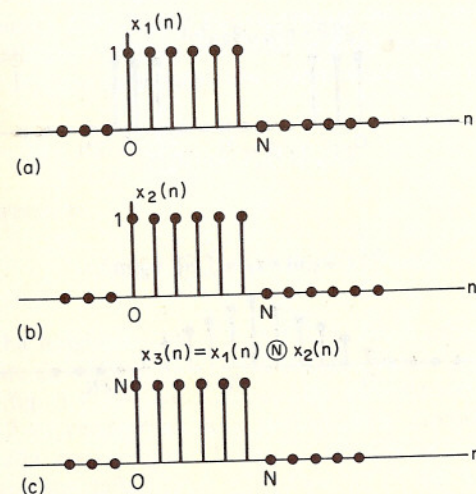


Fig. 3.11 Convoluzione circolare su N punti di due sequenze rettangolari di durata N .

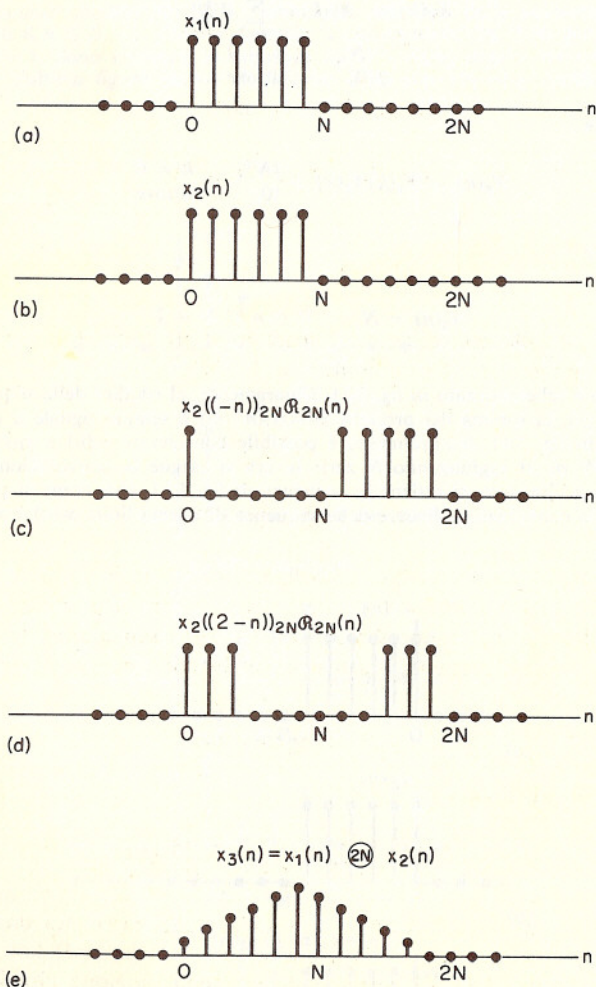


Fig. 3.12 Convoluzione circolare su $2N$ punti di due sequenze rettangolari di durata N .

Quest'ultimo esempio mette in evidenza un'utile interpretazione della convoluzione circolare. Si considerino due sequenze di durata finita $x_1(n)$ e $x_2(n)$ aventi trasformata di Fourier

$$X_1(e^{j\omega}) = \sum_{n=0}^{N-1} x_1(n) e^{-j\omega n}$$

$$X_2(e^{j\omega}) = \sum_{n=0}^{N-1} x_2(n) e^{-j\omega n}$$

La sequenza $x_3(n)$ che corrisponde al prodotto

$$X_3(e^{j\omega}) = X_1(e^{j\omega}) X_2(e^{j\omega})$$

è data da

$$x_3(n) = \sum_{m=0}^{N-1} x_1(m) x_2(n-m)$$

equivale, cioè, alla convoluzione lineare di $x_1(n)$ con $x_2(n)$. La sequenza risultante è lunga $2N-1$ campioni. Ora, le DFT

$$X_1(k) = \sum_{n=0}^{N-1} x_1(n) W_N^{nk}$$

$$X_2(k) = \sum_{n=0}^{N-1} x_2(n) W_N^{nk}$$

rappresentano i valori delle trasformate di Fourier $X_1(e^{j\omega})$ e $X_2(e^{j\omega})$ alle frequenze $\omega_k = 2\pi k/N$, che corrispondono ad un periodo di campionamento adeguato per rappresentare $x_1(n)$ e $x_2(n)$ senza *aliasing* nel dominio del tempo. La sequenza $x_4(n)$ che corrisponde alla relazione tra trasformate

$$X_4(k) = X_1(k) X_2(k)$$

è data dall'espressione

$$x_4(n) = \left[\sum_{r=-\infty}^{\infty} x_3(n+rN) \right] R_N(n) \quad (3.40)$$

Siccome $x_3(n)$ ha lunghezza $2N-1$, è chiaro che $x_4(n)$ sarà una versione della $x_3(n)$ affetta da *aliasing*. Ciò si può vedere confrontando le fig. 3.11 (c) e 3.12 (e): la fig. 3.11 (c) corrisponde alla convoluzione circolare su $2N$ punti e la 3.12 (e) corrisponde alla convoluzione circolare su $2N$ punti, che equivale alla convoluzione lineare delle due sequenze. Applicando l'operazione espressa dalla (3.40) alla sequenza di fig. 3.12 (e), si ottiene la sequenza di fig. 3.11 (c); si può vedere che succede proprio così, sommando la seconda metà della sequenza triangolare di fig. 3.12 (e) alla prima metà e moltiplicando il risultato per $R_N(n)$.

3.7 RIASSUNTO DELLE PROPRIETÀ DELLA TRASFORMATA DI FOURIER DISCRETA

Le proprietà della trasformata di Fourier discreta illustrate nei paragrafi precedenti sono riassunte nella tab. 3.2.

Tab. 3.2

Sequenza di lunghezza finita (N)	DFT
1. $x(n)$	$X(k)$
2. $y(n)$	$Y(k)$
3. $ax(n) + by(n)$	$aX(k) + bY(k)$
4. $x((n+m))_N \mathcal{R}_N(n)$	$W_N^{-km} X(k)$
5. $W_N^{ln} x(n)$	$X((k+l))_N \mathcal{R}_N(k)$
6. $\left[\sum_{m=0}^{N-1} x((m))_N y((n-m))_N \right] \mathcal{R}_N(n)$	$X(k) Y(k)$
7. $x(n)y(n)$	$\frac{1}{N} \left[\sum_{l=0}^{N-1} X((l))_N Y((k-l))_N \right] \mathcal{R}_N(k)$
8. $x^*(n)$	$X^*((-k))_N \mathcal{R}_N(k)$
9. $x^*((-n))_N \mathcal{R}_N(n)$	$X^*(k)$
10. $\text{Re} [x(n)]$	$X_{\text{ep}}(k) = \frac{1}{2} [X((k))_N + X^*((-k))_N] \mathcal{R}_N(k)$
11. $j \text{Im} [x(n)]$	$X_{\text{op}}(k) = \frac{1}{2} [X((k))_N - X^*((-k))_N] \mathcal{R}_N(k)$
12. $x_{\text{ep}}(n)$	$\text{Re} [X(k)]$
13. $x_{\text{op}}(n)$	$j \text{Im} [X(k)]$
Le proprietà seguenti valgono solo quando $x(n)$ è reale:	
14. $x(n)$ reale qualsiasi	$\begin{cases} X(k) = X^*((-k))_N \mathcal{R}_N(k) \\ \text{Re} [X(k)] = \text{Re} [X((k))_N] \mathcal{R}_N(k) \\ \text{Im} [X(k)] = -\text{Im} [X((k))_N] \mathcal{R}_N(k) \\ X(k) = X((k))_N \mathcal{R}_N(k) \\ \arg [X(k)] = -\arg [X((k))_N] \mathcal{R}_N(k) \end{cases}$
15. $x_{\text{ep}}(n)$	$\text{Re} [X(k)]$
16. $x_{\text{op}}(n)$	$j \text{Im} [X(k)]$

3.8 CONVOLUZIONE LINEARE BASATA SULLA TRASFORMATA DI FOURIER DISCRETA

Come mostrefemo in un successivo capitolo, esistono algoritmi molto efficienti per calcolare la trasformata di Fourier discreta di una sequenza di durata finita. Per questo motivo può convenire, dal punto di vista dei calcoli, effettuare la convoluzione di due sequenze calcolando le loro tra-

sformate di Fourier discrete, facendone il prodotto, e di questo calcolando poi la trasformata di Fourier discreta inversa. Nella maggior parte delle applicazioni interessa eseguire la convoluzione lineare di due sequenze. Questo è certamente vero, per esempio, quando si desidera filtrare una sequenza del tipo segnale voce o segnale radar. Come abbiamo visto nel precedente paragrafo, moltiplicare le trasformate di Fourier discrete di due sequenze corrisponde a calcolare la loro convoluzione circolare. Se quindi siamo interessati a ottenere una convoluzione lineare, dobbiamo assicurarci che la convoluzione circolare produca l'effetto di una convoluzione lineare. La chiave del metodo per ottenere questo risultato è messa in evidenza nel secondo esempio del par. 3.6.4.

Consideriamo innanzitutto due sequenze lunghe N , $x_1(n)$ e $x_2(n)$, e indichiamo con $x_3(n)$ la loro convoluzione lineare, cioè

$$x_3(n) = \sum_{m=0}^{N-1} x_1(m) x_2(n-m)$$

È immediato verificare che $x_3(n)$ è di lunghezza $2N-1$; ovvero essa può avere al più $2N-1$ punti diversi da zero. Se la si pensa ottenuta moltiplicando le trasformate discrete di Fourier di $x_1(n)$ ed $x_2(n)$, allora anche ciascuna di queste trasformate discrete, $X_1(k)$ e $X_2(k)$, deve essere stata calcolata sulla base di $2N-1$ punti. Pertanto definiamo

$$\begin{aligned} X_1(k) &= \sum_{n=0}^{2N-2} x_1(n) W_{2N-1}^{nk} \\ X_2(k) &= \sum_{n=0}^{2N-2} x_2(n) W_{2N-1}^{nk} \\ x_3(n) &= \frac{1}{2N-1} \left[\sum_{k=0}^{2N-2} [X_1(k) X_2(k)] W_{2N-1}^{-nk} \right] \mathcal{R}_{2N-1}(n) \end{aligned} \quad (3.41)$$

e ne deduciamo che $x_3(n)$ sarà la convoluzione lineare di $x_1(n)$ e $x_2(n)$. Ovviamente otterremmo una convoluzione lineare anche se le trasformate di Fourier discrete fossero calcolate sulla base di più di $2N-1$ punti, ma non la otterremmo, in generale, se le DFT fossero calcolate sulla base di un numero più piccolo di punti. Un altro modo di considerare questo procedimento che consente di ottenere la convoluzione lineare, consiste nel notare che il calcolo delle DFT sulla base di $2N-1$ punti corrisponde a delle serie di Fourier per sequenze periodiche costruite da $x_1(n)$ e $x_2(n)$ in modo tale che gli ultimi $N-1$ punti in ogni periodo sono zero. Queste sequenze periodiche sono illustrate nella fig. 3.13. Questa figura mostra anche il modo di ottenere la convoluzione periodica e consente di notare che, a causa degli zeri aggiunti in ogni periodo, i

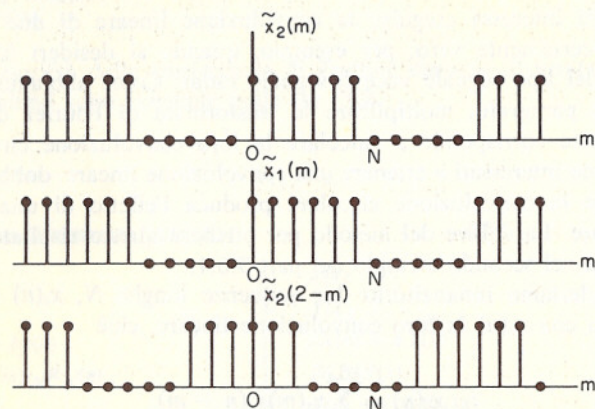


Fig. 3.13 Sequenze periodiche di periodo $(2N - 1)$ costruite da sequenze finite di durata N . Gli ultimi $(N - 1)$ punti in ogni periodo sono zero.

valori diversi da zero in un periodo di $\tilde{x}_i(n)$ vengono interessati soltanto dai valori diversi da zero in un singolo periodo di $\tilde{x}_i(n)$.

In generale si può voler fare la convoluzione di due sequenze di durata diversa. Se $x_1(n)$ ha durata N_1 ed $x_2(n)$ ha durata N_2 , allora la loro convoluzione sarà lunga $N_1 + N_2 - 1$. Pertanto in questo caso andranno moltiplicate tra loro le trasformate di Fourier discrete calcolate sulla base di $N \geq N_1 + N_2 - 1$.

Il procedimento appena descritto consente di calcolare la convoluzione lineare di due sequenze di durata finita facendo uso della trasformata di Fourier discreta. In alcune applicazioni si vorrebbe tuttavia fare la convoluzione di una sequenza di durata finita con una sequenza di durata infinita, come, per esempio, nel caso del filtraggio della voce. In via teorica noi potremmo registrare l'intero segnale e poi attuare il procedimento visto sopra sulla base di una DFT di un gran numero di punti; ma le dimensioni di una tale DFT ne rendono generalmente impossibile il calcolo. Un'altra considerazione da fare è che con questo metodo non si potrebbe calcolare alcun punto della sequenza filtrata prima di aver raccolto tutti i punti della sequenza di ingresso. Generalmente si desidera invece evitare simili ritardi nella elaborazione. Per ottenere questo risultato pur continuando ad usare la trasformata di Fourier discreta, il segnale da filtrare deve essere segmentato in sezioni di lunghezza L [3,4]. Ogni segmento può quindi essere convoluto con la risposta all'impulso di durata finita e i segmenti filtrati congiunti infine uno all'altro in modo opportuno. Una simile tecnica di filtraggio a blocchi può pertanto essere attuata usando, come prima, la trasformata di Fourier discreta.

Per illustrare il procedimento e mostrare al tempo stesso il modo di congiungere l'uno all'altro i vari segmenti filtrati, si considerino la risposta all'impulso $h(n)$ di lunghezza M e il segnale $x(n)$ raffigurati in fig. 3.14. Scomponiamo $x(n)$ in una somma di segmenti ciascuno avente soltanto L punti diversi da zero, e indichiamo il k -mo segmento con $x_k(n)$, dove

$$x_k(n) = \begin{cases} x(n), & kL \leq n \leq (k+1)L - 1 \\ 0, & \text{altrove} \end{cases} \quad (3.42)$$

Pertanto $x(n)$ è la somma delle $x_k(n)$, cioè

$$x(n) = \sum_{k=0}^{\infty} x_k(n) \quad (3.43)$$

e la convoluzione di $x(n)$ con $h(n)$ è uguale alla somma delle convoluzioni delle $x_k(n)$ con $h(n)$, ovvero

$$x(n) * h(n) = \sum_{k=0}^{\infty} x_k(n) * h(n) \quad (3.44)$$

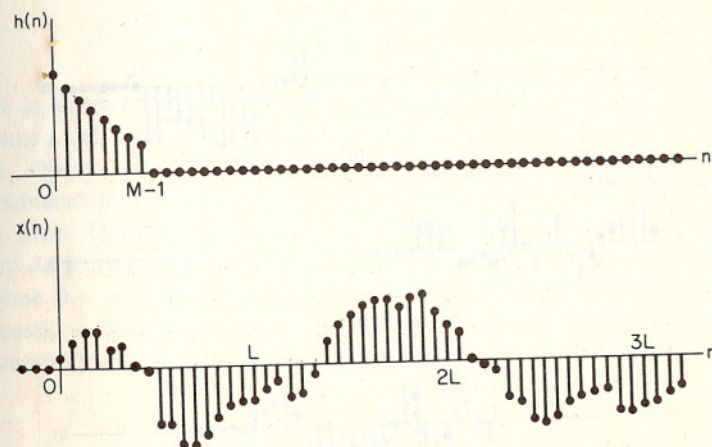


Fig. 3.14 Risposta all'impulso $h(n)$ di durata finita e segnale $x(n)$ da filtrare.

Poiché le $x_k(n)$ hanno solo L punti diversi da zero e $h(n)$ è di lunghezza M , ciascuno dei termini $[x_k(n) * h(n)]$ nella somma è di lunghezza $L + M - 1$. Pertanto la convoluzione lineare $x(n) * h(n)$ può essere ottenuta usando una DFT di $(L + M - 1)$ punti. Poiché inoltre la distanza tra i campioni iniziali di due segmenti di ingresso adiacenti è di L punti, e ogni segmento filtrato ha lunghezza $(L + M - 1)$, ne segue che i punti diversi da zero nei segmenti filtrati si sovrappongono, nello svolgimento della (3.44), di $(M - 1)$ punti. Ciò è illustrato nella fig. 3.15. Nelle

fig. 3.15 (a) e 3.15 (b) sono raffigurati rispettivamente i segmenti di ingresso $x_k(n)$ e i segmenti filtrati $x_k(n)*h(n)$. L'intera forma d'onda di ingresso, $x(n)$, si ricostruisce sommando le forme d'onda di fig. 3.15 (a), e il risultato filtrato, $x(n)*h(n)$, si costruisce sommando i segmenti filtrati

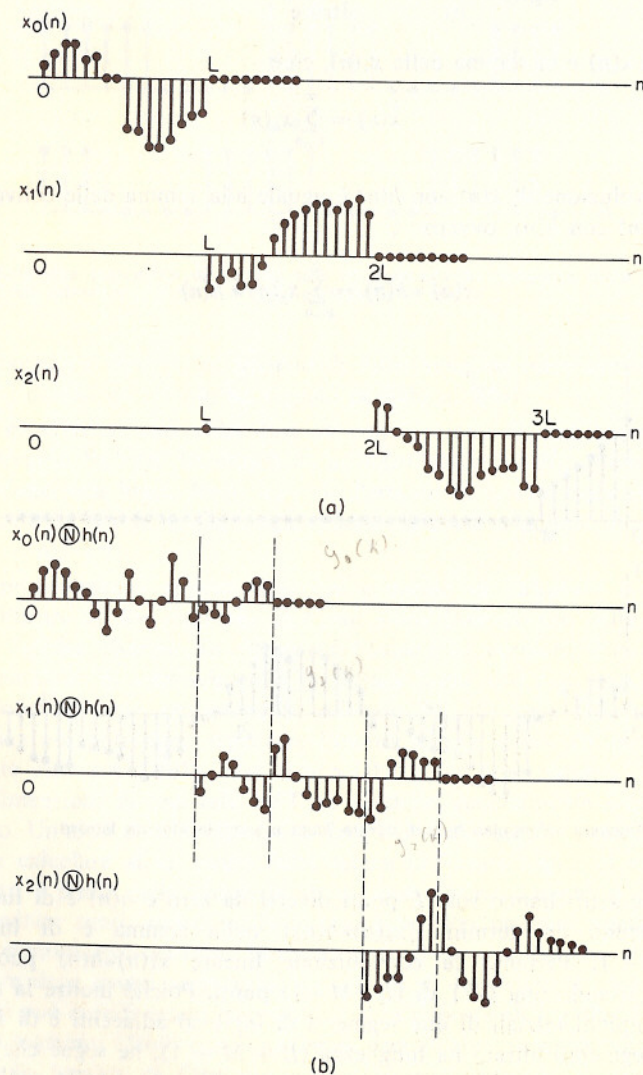


Fig. 3.15 (a) Scomposizione di $x(n)$ in segmenti non sovrappoventisi di lunghezza L ; (b) risultato della convoluzione di ogni segmento con $h(n)$.

raffigurati in fig. 3.15 (b). Questo procedimento di costruzione dell'uscita filtrata è spesso chiamato il *metodo di sovrapposizione e somma*, in relazione al fatto che i segmenti filtrati sono sovrapposti e sommati per costruire l'uscita. La sovrapposizione deriva dal fatto che la convoluzione lineare di ogni segmento con la risposta all'impulso è in generale più lunga del segmento stesso.

Un procedimento alternativo, comunemente chiamato *metodo di sovrapposizione ed estrazione*, consiste nel calcolare una convoluzione circolare fra $h(n)$ ed $x_k(n)$, identificando poi quella parte della convoluzione circolare che corrisponde a una convoluzione lineare. In particolare, se consideriamo la convoluzione circolare della risposta all'impulso lunga M con un segmento lungo N , risulta che i primi $M - 1$ punti di tale convoluzione non sono corretti, mentre i rimanenti punti sono gli stessi che otterremmo dalla convoluzione lineare. In questo caso, quindi, conviene sezionare $x(n)$ in segmenti di lunghezza N in modo tale che ogni segmento di ingresso si sovrapponga al precedente per $M - 1$ punti. Definiamo perciò i segmenti $x_k(n)$ come

$$x_k(n) = x(n + k(N - M + 1)), \quad 0 \leq n \leq N - 1$$

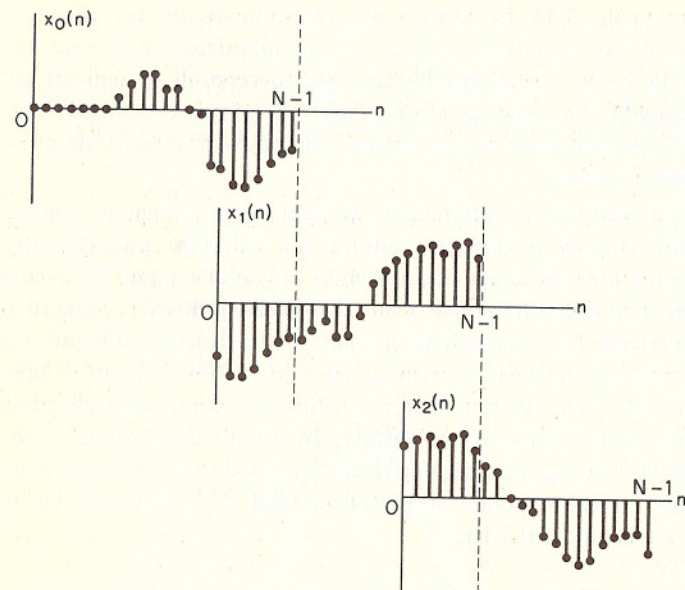
dove in questo caso abbiamo posto l'origine temporale di ogni segmento all'inizio del segmento stesso piuttosto che coincidente con l'origine di $x(n)$. Questo metodo di segmentazione è raffigurato nella fig. 3.16 (a). Indichiamo inoltre con $y'_k(n)$ le convoluzioni circolari di ogni segmento con $h(n)$. Queste convoluzioni sono raffigurate nella fig. 3.16 (b). La parte da scartare per ogni sequenza di uscita è quella che corrisponde alla regione $0 \leq n \leq M - 2$. I punti che restano dalle sequenze di uscita in successione devono poi essere « giuntati » gli uni agli altri in modo da ottenere l'uscita filtrata finale. Si ha quindi

$$y(n) = \sum_{k=0}^{\infty} y'_k(n - k(N + M - 1))$$

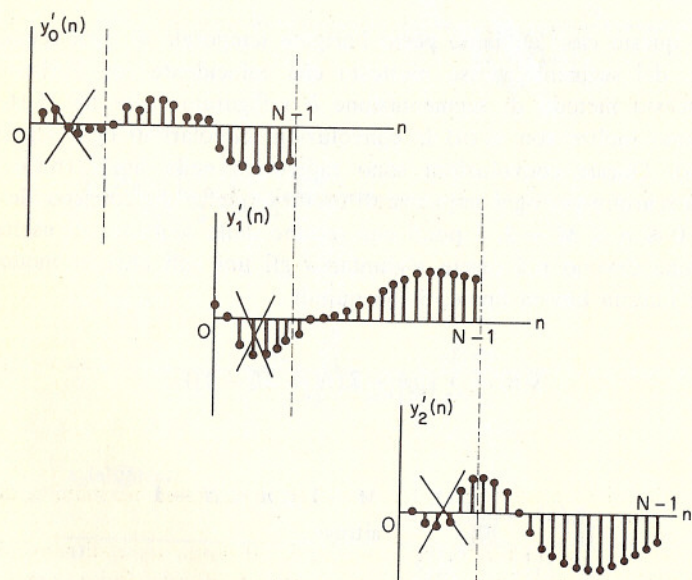
dove

$$y'_k(n) = \begin{cases} y'_k(n), & M - 1 \leq n \leq N - 1 \\ 0, & \text{altrove} \end{cases}$$

Questo procedimento, ovvero il metodo di sovrapposizione ed estrazione, deve il suo nome al fatto che ogni successivo segmento di ingresso consiste di $N - M + 1$ nuovi punti e di $M - 1$ punti conservati dal segmento precedente.



(a)



(b)

Fig. 3.16 (a) Scomposizione di $x(n)$ in segmenti sovrapponibili di lunghezza N ; (b) risultato della convoluzione circolare di ogni segmento con $h(n)$. Sono anche indicate le porzioni da scartare per ogni segmento filtrato allo scopo di ottenere la convoluzione lineare.

3.9 TRASFORMATTA DI FOURIER DISCRETA BIDIMENSIONALE

Nei primi due capitoli abbiamo visto come molte delle proprietà delle trasformate valide per segnali a una dimensione possono essere estese a segnali a più dimensioni. Una generalizzazione simile vale per la serie e la trasformata di Fourier discreta.

La rappresentazione di sequenze bidimensionali mediante la trasformata di Fourier discreta ha notevole importanza nella elaborazione numerica di segnali bidimensionali come fotografie o dati sismici. In questo paragrafo ci limiteremo a una breve discussione delle DFS e DFT bidimensionali seguendo le linee della discussione svolta nei precedenti paragrafi di questo capitolo.

Cominciamo considerando la definizione di una sequenza periodica bidimensionale. Diremo che una sequenza è periodica nell'indice delle righe con periodo M e nell'indice delle colonne con periodo N se è

$$\tilde{x}(m, n) = \tilde{x}(m + qM, n + rN)$$

dove q ed r sono numeri interi arbitrari positivi o negativi. Tali sequenze hanno una rappresentazione mediante serie di Fourier come somma di esponenziali complessi nella forma

$$\tilde{x}(m, n) = \frac{1}{MN} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \tilde{X}(k, l) W_M^{-km} W_N^{-ln} \quad (3.45)$$

dove $\tilde{X}(k, l)$, come si può dimostrare, vale

$$\tilde{X}(k, l) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \tilde{x}(m, n) W_M^{km} W_N^{ln} \quad (3.46)$$

con

$$W_M = e^{-j(2\pi/M)}$$

$$W_N = e^{-j(2\pi/N)}$$

Dalla (3.46) si può verificare che

$$\tilde{X}(k, l) = \tilde{X}(k + qM, l + rN)$$

per valori interi di q ed r , e pertanto $\tilde{X}(k, l)$ ha la stessa periodicità della sequenza $\tilde{x}(m, n)$.

Si è visto come la trasformata di Fourier discreta monodimensionale deriva dalla interpretazione di una sequenza di durata finita come un periodo di una sequenza periodica cui si applica la serie di Fourier discreta. Analogamente si può applicare la serie di Fourier bidimensionale per rappresentare una sequenza bidimensionale che è diversa da zero solo in una regione finita del piano (m, n) . Una tale sequenza verrà chia-

mata sequenza ad area finita ed è il corrispettivo bidimensionale delle sequenze di durata finita. La rappresentazione di Fourier che ne risulta viene chiamata trasformata di Fourier discreta bidimensionale.

Onde sviluppare la DFT per segnali bidimensionali consideriamo una sequenza ad area finita $x(m, n)$ che è zero al di fuori dell'intervallo $0 \leq m \leq M-1$, $0 \leq n \leq N-1$, e quindi di area (M, N) , e costruiamo la sequenza periodica

$$\tilde{x}(m, n) = x[((m))_M, ((n))_N] \quad (3.47)$$

La sequenza originaria $x(m, n)$ si riottiene estraendo un singolo periodo di $\tilde{x}(m, n)$, ovvero

$$x(m, n) = \tilde{x}(m, n) \mathcal{R}_{M,N}(m, n) \quad (3.48)$$

dove

$$\mathcal{R}_{M,N}(m, n) = \begin{cases} 1, & 0 \leq m \leq M-1, 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (3.49)$$

Stabiliamo ora che la trasformata di Fourier discreta di $x(m, n)$ corrisponde ai coefficienti della serie di Fourier di $\tilde{x}(m, n)$. Tuttavia, proprio come abbiamo fatto per le sequenze monodimensionali, noi interpreteremo i coefficienti della DFT come una sequenza ad area finita per mantenere anche in questo caso la dualità fra il dominio delle frequenze e quello originario. Pertanto, indicando con $X(k, l)$ la DFT di $x(m, n)$, si ha

$$X(k, l) = \left[\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m, n) W_M^{km} W_N^{ln} \right] \mathcal{R}_{M,N}(k, l) \quad (3.50)$$

$$x(m, n) = \frac{1}{MN} \left[\sum_{k=0}^{M-1} \sum_{l=0}^{N-1} X(k, l) W_M^{-km} W_N^{-ln} \right] \mathcal{R}_{M,N}(m, n) \quad (3.51)$$

Si può mettere in evidenza un'utile interpretazione della DFT bidimensionale in termini di quella monodimensionale osservando che la funzione rettangolare $\mathcal{R}_{M,N}(k, l)$ è separabile e pertanto può essere scritta come

$$\mathcal{R}_{M,N}(k, l) = \mathcal{R}_M(k) \mathcal{R}_N(l) \quad (3.52)$$

Di conseguenza la (3.50) può scriversi

$$X(k, l) = \left[\sum_{n=0}^{N-1} G(k, n) W_N^{ln} \right] \mathcal{R}_N(l) \quad (3.53a)$$

dove

$$G(k, n) = \left[\sum_{m=0}^{M-1} x(m, n) W_M^{km} \right] \mathcal{R}_M(k) \quad (3.53b)$$

La funzione $G(k, n)$ corrisponde, per ogni valore di n , a una DFT monodimensionale a M punti; essa è costituita cioè da N trasformate monodimensionali, una per ogni colonna di $x(m, n)$. La DFT bidimensionale $X(k, l)$ si ottiene allora, in accordo con la (3.53 a), effettuando M trasformate monodimensionali, una per ogni riga della sequenza $G(k, n)$.

La (3.50) può alternativamente scriversi come

$$X(k, l) = \left[\sum_{m=0}^{M-1} P(m, l) W_M^{km} \right] \mathcal{R}_M(k) \quad (3.54a)$$

dove

$$P(m, l) = \left[\sum_{n=0}^{N-1} x(m, n) W_N^{ln} \right] \mathcal{R}_N(l) \quad (3.54b)$$

La funzione $P(m, l)$ corrisponde ora a un insieme di trasformate a N punti sulle righe di $x(m, n)$. La $X(k, l)$ si ottiene allora, in accordo con la (3.54 a), trasformando le colonne di $P(m, l)$. In sintesi, dunque, le DFT bidimensionali si possono calcolare usando trasformate monodimensionali prima sulle righe e poi sulle colonne, o viceversa. Un ragionamento analogo può, ovviamente, farsi anche per la DFT inversa rappresentata dalla (3.51).

Un caso di particolare interesse è quello delle sequenze separabili, aventi cioè la proprietà che

$$x(m, n) = x_1(m) x_2(n) \quad (3.55)$$

In questo caso la funzione $G(k, n)$ nella (3.53 b) è $X_1(k)$, cioè la DFT monodimensionale di $x_1(m)$, ed è indipendente da n . La DFT bidimensionale è il prodotto di $X_1(k)$ e $X_2(l)$, la DFT di $x_2(n)$, e cioè

$$X(k, l) = X_1(k) X_2(l) \quad (3.56)$$

In questo caso il solo calcolo di una DFT a M punti e di una DFT a N punti ci permette di calcolare $X(k, l)$ per tutti i k ed l .

La trasformata di Fourier discreta bidimensionale è chiaramente lineare, ovvero, se

$$x_3(m, n) = ax_1(m, n) + bx_2(m, n)$$

allora

$$X_3(k, l) = aX_1(k, l) + bX_2(k, l)$$

dove si è fatta l'ipotesi che $x_1(m, n)$ e $x_2(m, n)$ abbiano dimensioni identiche.

Nell'ambito delle sequenze monodimensionali di durata finita abbiamo notato come una traslazione nel dominio temporale possa essere interpretata come una rotazione nell'intervallo base $0 \leq n \leq N-1$. Nel

caso della sequenza bidimensionale di area finita $x(m + m_0, n + n_0)$, si può dimostrare come la corrispondente DFT sia $W_M^{-km_0} W_N^{-ln_0} X(k, l)$. In questo caso possiamo interpretare la traslazione nel dominio spaziale come l'operazione di ruotare di m_0 campioni ogni colonna, seguita da una rotazione di n_0 campioni di ogni riga della nuova sequenza bidimensionale. Questa proprietà, ovviamente, ha una formulazione simmetrica quando vengono scambiati i domini spazio e frequenza.

Come per la DFT monodimensionale, esiste tutto un insieme di proprietà di simmetria per la DFT bidimensionale. Alcune di queste sono discusse nel probl. 35 di questo capitolo.

Un'applicazione importante della DFT bidimensionale riguarda il calcolo di convoluzioni a scopo di filtraggio. Si considerino due sequenze di area finita $x_1(m, n)$ e $x_2(m, n)$, dove $x_1(m, n)$ è di area (M_1, N_1) e $x_2(m, n)$ è di area (M_2, N_2) . Indichiamo con $X_1(k, l)$ e $X_2(k, l)$ le DFT di dimensioni (M, N) rispettivamente di $x_1(m, n)$ e $x_2(m, n)$, aumentate se necessario di aree di campioni nulli. Si ha che il prodotto

$$X_3(k, l) = X_1(k, l) X_2(k, l) \quad (3.57)$$

corrisponde alla sequenza

$$x_3(m, n) = \sum_{q=0}^{M-1} \sum_{r=0}^{N-1} x_1[(q)]_M [(r)]_N x_2[(m-q)]_M [(n-r)]_N \mathcal{R}_{M,N}(m, n) \quad (3.58)$$

L'espressione (3.58) rappresenta la convoluzione periodica delle sequenze periodiche $\tilde{x}_1(m, n)$ e $\tilde{x}_2(m, n)$ formate da $x_1(m, n)$ e $x_2(m, n)$ come in (3.47). Nel contesto delle sequenze ad area finita, la (3.58) è una convoluzione circolare in due dimensioni. Se quello che desideriamo ottenere è la convoluzione lineare di $x_1(m, n)$ e $x_2(m, n)$, dobbiamo ovviamente accertarci che, come nel caso monodimensionale, M ed N siano scelti in modo da evitare *aliasing*. Poiché la convoluzione di una sequenza che ricopre l'area (M_1, N_1) con una sequenza che ricopre l'area (M_2, N_2) , dà luogo a una sequenza di area $[(M_1 + M_2 - 1), (N_1 + N_2 - 1)]$, dobbiamo scegliere $M \geq M_1 + M_2 - 1$ ed $N \geq N_1 + N_2 - 1$ se vogliamo essere sicuri che la convoluzione circolare sia identica alla convoluzione lineare desiderata.

Tutto ciò è illustrato nella fig. 3.17, dove le regioni diverse da zero di $x_1(m, n)$ e $x_2(m, n)$ risultano tratteggiate. In questa figura abbiamo sovrapposto $x_2(m - q, n - r)$ a $x_1(q, r)$. Chiaramente, se le disuguaglianze scritte sopra risultano soddisfatte, $x_2(m - q, n - r)$ nella sua parte « riavvolta » non verrà mai ad interessare porzioni di $x_1(q, r)$ diverse da zero, e pertanto la convoluzione circolare sarà identica alla convoluzione lineare desiderata. Se si dovesse fare la convoluzione di un'area piccola

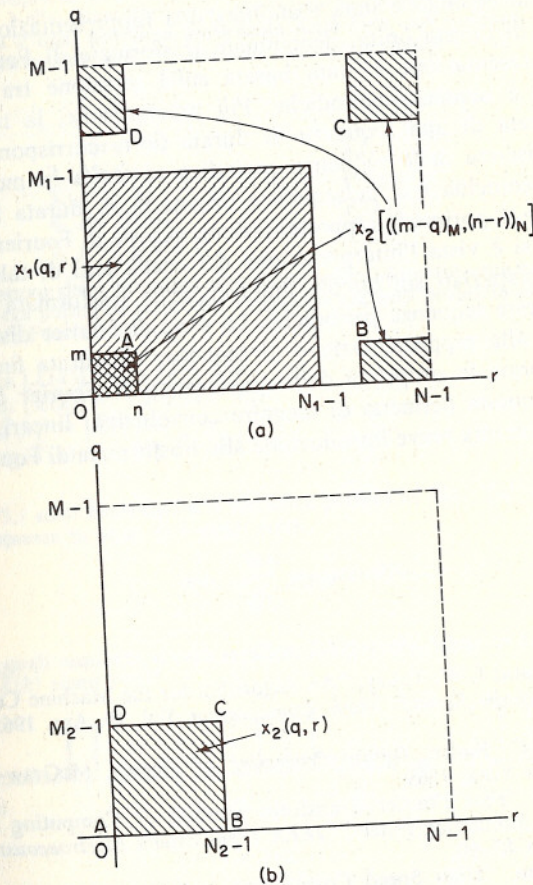


Fig. 3.17 Realizzazione di una convoluzione lineare bidimensionale per mezzo di una convoluzione circolare: (a) $x_1(q, r)$ e $x_2[(m-q)]_M [(n-r)]_N$; (b) $x_2(q, r)$.

con un'area molto più grande, potremmo generalizzare i metodi di sovrapposizione e somma e di sovrapposizione ed estrazione considerati nel paragrafo precedente. Inoltre, se una delle sequenze da convolvere è separabile, la convoluzione bidimensionale può effettuarsi mediante il calcolo ripetuto di convoluzioni monodimensionali. Sequenze separabili sono spesso usate nelle applicazioni di filtraggio, proprio grazie alle semplificazioni di calcolo che consentono e che sono state discusse in precedenza.

SOMMARIO

In questo capitolo è stata esaminata una rappresentazione di Fourier di sequenze di durata finita, denominata trasformata di Fourier discreta. Questa rappresentazione è stata basata sulla relazione tra sequenze di durata finita e sequenze periodiche. Più precisamente, la trasformata di Fourier discreta di una sequenza di durata finita corrisponde alla serie di Fourier discreta della sequenza periodica costruita in modo che ogni suo periodo coincida con la sequenza originaria di durata finita. Perciò è stata trattata dapprima la rappresentazione in serie di Fourier di sequenze periodiche e si è vista l'interpretazione dei coefficienti di tale serie come campioni, equispaziati sul circolo unitario, della trasformata z di un periodo delle stesse sequenze periodiche. La serie di Fourier discreta è stata poi applicata alla rappresentazione di sequenze di durata finita. Si sono anche considerate le proprietà della trasformata di Fourier discreta e si è visto come questa permetta di eseguire convoluzioni lineari. Il capitolo si è concluso con una breve introduzione alla trasformata di Fourier discreta bidimensionale.

BIBLIOGRAFIA

1. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Computation of Complex Fourier Series," *Math. Computation*, Vol. 19, Apr. 1965, pp. 297-301.
2. B. Gold and C. Rader, *Digital Processing of Signals*, McGraw-Hill Book Company, New York, 1969.
3. H. D. Helms, "Fast Fourier Transform Method of Computing Difference Equations and Simulating Filters," *IEEE Trans. Audio Electroacoust.*, Vol. 15, No. 2, 1967, pp. 85-90.
4. T. G. Stockham, "High Speed Convolution and Correlation," 1966 Spring Joint Computer Conference, *AFIPS Proc.*, Vol. 28, 1966, pp. 229-233.

PROBLEMI

1. Si consideri il sistema lineare del primo ordine definito dall'equazione alle differenze $y(n) = ay(n-1) + x(n)$, in cui il coefficiente a è compreso tra zero e l'unità. L'ingresso $x(n]$ è vincolato ad essere una sequenza periodica di periodo N , cioè $\tilde{x}(n) = \tilde{x}(n + kN)$ per ogni k intero. Si assume che l'uscita del filtro abbia raggiunto il regime. Determinare, in termini del coefficiente a , la risposta all'impulso di un filtro a risposta all'impulso finita che fornisca, per questa classe di ingressi, un'uscita $\tilde{y}(n)$ indistinguibile, a regime, da quella del filtro a risposta all'impulso infinita definito dalla precedente equazione alle differenze.

2. (a) Nel par. 3.3.3 abbiamo enunciato un certo numero di proprietà di simmetria della serie di Fourier discreta per sequenze periodiche. Elenchiamo qui alcune delle proprietà enunciate. Dimostrate che ciascuna delle proprietà elencate è vera. Nella dimostrazione si può usare la definizione della serie di Fourier discreta e ogni proprietà precedente nella lista. Ad esempio, nella dimostrazione della proprietà 3 si possono usare le proprietà 1 e 2.

Sequenza	Serie di Fourier discreta
1. $\tilde{x}(n + m)$	$W_N^{-km} \tilde{X}(k)$
2. $\tilde{x}^*(n)$	$\tilde{X}^*(-k)$
3. $\tilde{x}^*(-n)$	$\tilde{X}^*(k)$
4. $\text{Re} [\tilde{x}(n)]$	$\tilde{X}_e(k)$
5. $j \text{Im} [\tilde{x}(n)]$	$\tilde{X}_o(k)$

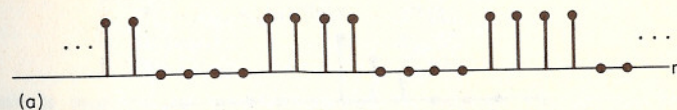
- (b) Per mezzo delle proprietà dimostrate nella parte (a), mostrare che per una sequenza periodica reale $\tilde{x}(n)$ valgono le seguenti proprietà di simmetria della serie di Fourier discreta:

- (1) $\text{Re} [\tilde{X}(k)] = \text{Re} [\tilde{X}(-k)]$.
- (2) $\text{Im} [\tilde{X}(k)] = -\text{Im} [\tilde{X}(-k)]$.
- (3) $|\tilde{X}(k)| = |\tilde{X}(-k)|$.
- (4) $\arg \tilde{X}(k) = -\arg \tilde{X}(-k)$.

3. In fig. P3.3 sono mostrate varie sequenze periodiche $\tilde{x}(n)$. Tali sequenze possono essere espresse in serie di Fourier come

$$\tilde{x}(n) = \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)kn}$$

- (a) Per quali sequenze si può scegliere l'origine dei tempi in modo tale che tutti gli $\tilde{X}(k)$ siano reali?



- (b) Per quali sequenze si può scegliere l'origine dei tempi in modo tale che tutti gli $\tilde{X}(k)$ (escluso $\tilde{X}(0)$) siano immaginari?
- (c) Per quali sequenze è $\tilde{X}(k) = 0$, $k = \pm 2, \pm 4, \pm 6$, ecc.?
4. Se $\tilde{x}(n)$ è una sequenza periodica con periodo N , è anche periodica con periodo $2N$. Si indichino con $\tilde{X}_1(k)$ i coefficienti della DFS di $\tilde{x}(n)$ considerata come sequenza periodica con periodo N e con $\tilde{X}_2(k)$ i coefficienti della DFS di $\tilde{x}(n)$ considerata come sequenza periodica con periodo $2N$. $\tilde{X}_1(k)$ è ovviamente periodica con periodo N e $\tilde{X}_2(k)$ è periodica con periodo $2N$. Determinare $\tilde{X}_2(k)$ in funzione di $\tilde{X}_1(k)$.
5. Si considerino due sequenze periodiche $\tilde{x}(n)$ e $\tilde{y}(n)$. $\tilde{x}(n)$ ha periodo N e $\tilde{y}(n)$ ha periodo M . La sequenza $\tilde{w}(n)$ è definita come $\tilde{w}(n) = \tilde{x}(n) + \tilde{y}(n)$.
- (a) Mostrare che $\tilde{w}(n)$ è periodica con periodo MN .
- (b) Poiché $\tilde{x}(n)$ ha periodo N , i coefficienti $\tilde{X}(k)$ della sua DFS hanno anch'essi periodo N . Analogamente, poiché $\tilde{y}(n)$ ha periodo M , i coefficienti $\tilde{Y}(k)$ della sua DFS hanno anch'essi periodo M . I coefficienti della DFS di $\tilde{w}(n)$, $\tilde{W}(k)$, hanno periodo MN . Determinare $\tilde{W}(k)$ in termini di $\tilde{X}(k)$ e $\tilde{Y}(k)$. Può essere utile far riferimento ai risultati del precedente probl. 4.
6. $\tilde{x}(n)$ indica una sequenza periodica con periodo N e $\tilde{X}(k)$ indica i coefficienti della sua serie di Fourier discreta. La sequenza $\tilde{X}(k)$ è anch'essa una sequenza periodica con periodo N . Determinare, in termini di $\tilde{x}(n)$, i coefficienti della serie di Fourier discreta di $\tilde{X}(k)$.
7. Calcolare la DFT di ciascuna delle seguenti sequenze di durata finita considerate di durata N .
- (a) $x(n) = \delta(n)$.
- (b) $x(n) = \delta(n - n_0)$, dove $0 < n_0 < N$.
- (c) $x(n) = a^n$, $0 \leq n \leq N - 1$.

8. In fig. P3.8 è mostrata una sequenza di durata finita $x(n)$. Rappresentare la sequenza $x((-n))_N$.

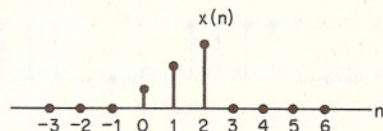


Fig. P3.8

9. $x(n)$ indica una sequenza di durata finita N . Mostrare che

$$x((-n))_N = x((N - n))_N$$

10. Dei dati analogici di cui si vuole analizzare lo spettro sono campionati a 10 kHz e viene calcolata la DFT di 1024 campioni. Determinare la distanza in frequenza tra i campioni dello spettro. Giustificare la risposta.
11. La DFT di una sequenza di durata finita corrisponde a campioni della sua trasformata z sul circolo unitario. Ad esempio, la DFT di una sequenza $x(n)$ di 10 punti corrisponde ai valori di $X(z)$ nei 10 punti equispaziati indicati in fig. P3.11-1. Vogliamo determinare i valori equispaziati di $X(z)$ sul contorno mostrato in fig. P3.11-2, cioè $X(z)|_{z=0.5e^{j[(2\pi k/10) + (\pi/10)]}}$. Mostrare come va modificata $x(n)$ per ottenere una sequenza $x_1(n)$ tale che la DFT di $x_1(n)$ corrisponda ai valori desiderati di $X(z)$.

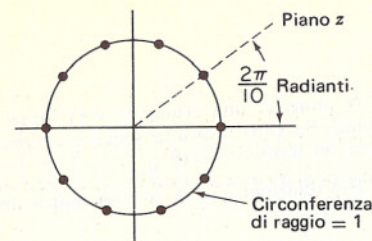


Fig. P3.11-1

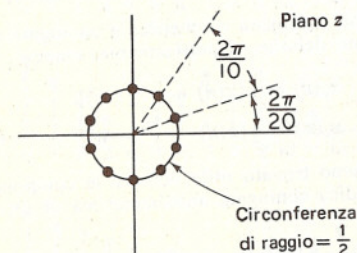


Fig. P3.11-2

12. Nel par. 3.7.3 abbiamo enunciato un certo numero di proprietà di simmetria della DFT, alcune delle quali sono qui elencate. Dimostrare che ciascuna delle proprietà elencate è vera. Nella dimostrazione si può usare la definizione della DFT e proprietà precedenti nella lista.

Sequenza	DFT
1. $x((n + m))_N \mathcal{R}_N(n)$	$W_N^{-km} X(k)$
2. $x^*(n)$	$X^*((-k))_N \mathcal{R}_N(k)$
3. $x^*((-n))_N \mathcal{R}_N(n)$	$X^*(k)$
4. $\text{Re}[x(n)]$	$X_{\text{ep}}(k)$
5. $j \text{Im}[x(n)]$	$X_{\text{op}}(k)$

13. Usando le proprietà del precedente probl. 12, mostrare che per una sequenza reale sono valide le seguenti proprietà di simmetria della DFT:

- (1) $\text{Re}[X(k)] = \text{Re}[X((-k))_N \mathcal{R}_N(k)]$.
- (2) $\text{Im}[X(k)] = -\text{Im}[X((-k))_N \mathcal{R}_N(k)]$.
- (3) $|X(k)| = |X((-k))_N \mathcal{R}_N(k)|$.
- (4) $\arg[X(k)] = -\arg[X((-k))_N \mathcal{R}_N(k)]$.

14. Sia $X(k)$ la DFT su N punti della sequenza $x(n)$ lunga N .

- (a) Mostrare che se $x(n)$ soddisfa la relazione

$$x(n) = -x(N - 1 - n)$$

allora è

$$X(0) = 0$$

- (b) Mostrare che con N pari e se

$$x(n) = x(N - 1 - n)$$

allora

$$X\left(\frac{N}{2}\right) = 0$$



15. Sia $X(k)$ la DFT su N punti di una sequenza $x(n)$ lunga N . $X(k)$ è essa stessa una sequenza di N punti. Se $x_1(n)$ indica la sequenza ottenuta eseguendo la DFT di $X(k)$, esprimere $x_1(n)$ in termini di $x(n)$.
16. Mostrare che in base alla relazione (3.26), se $x(n)$ indica una sequenza lunga N e $X(k)$ la sua DFT su N punti,

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2$$

Questo risultato è comunemente indicato come *relazione di Parseval per la DFT*.

17. Nel cap. 1 le componenti coniugata simmetrica e coniugata antisimmetrica di una sequenza $x(n)$ sono state definite, rispettivamente, come

$$x_e(n) = \frac{1}{2} [x(n) + x^*(-n)]$$

$$x_o(n) = \frac{1}{2} [x(n) - x^*(-n)]$$

Nel par. 3.6.3 abbiamo trovato utile definire le componenti periodica coniugata simmetrica e periodica coniugata antisimmetrica di una sequenza di durata finita N come

$$x_{ep}(n) = \frac{1}{2} [x((n))_N + x^*((-n))_N] \mathcal{R}_N(n)$$

$$x_{op}(n) = \frac{1}{2} [x((n))_N - x^*((-n))_N] \mathcal{R}_N(n)$$

- (a) Mostrare che $x_{ep}(n)$ può essere messa in relazione con $x_e(n)$ e $x_{op}(n)$ può essere messa in relazione con $x_o(n)$ per mezzo delle espressioni

$$x_{ep}(n) = [x_e(n) + x_e(n - N)] \mathcal{R}_N(n)$$

$$x_{op}(n) = [x_o(n) + x_o(n - N)] \mathcal{R}_N(n)$$

- (b) $x(n)$ è considerata una sequenza di durata N , e in generale $x_e(n)$ non può essere ricostruita da $x_{ep}(n)$ e $x_o(n)$ non può essere ricostruita da $x_{op}(n)$. Mostrare che con $x(n)$ considerata di durata N , ma con $x(n) = 0$ per $n > N/2$, $x_e(n)$ può essere ottenuta da $x_{ep}(n)$ e $x_o(n)$ può essere ottenuta da $x_{op}(n)$.

18. Una sequenza $x(n)$ di durata finita lunga 8 ha la DFT sulla base di 8 punti $X(k)$ mostrata in fig. P3.18-1. Una nuova sequenza $y(n)$ di durata 16 è definita da

$$y(n) = \begin{cases} x\left(\frac{n}{2}\right), & n \text{ pari} \\ 0, & n \text{ dispari} \end{cases}$$

Dalla lista in fig. P3.18-2, scegliere il grafico corrispondente alla DFT su 16 punti di $y(n)$.

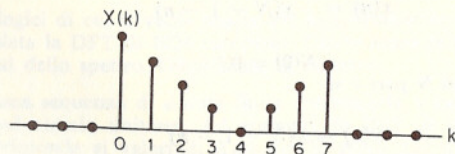


Fig. P3.18-1

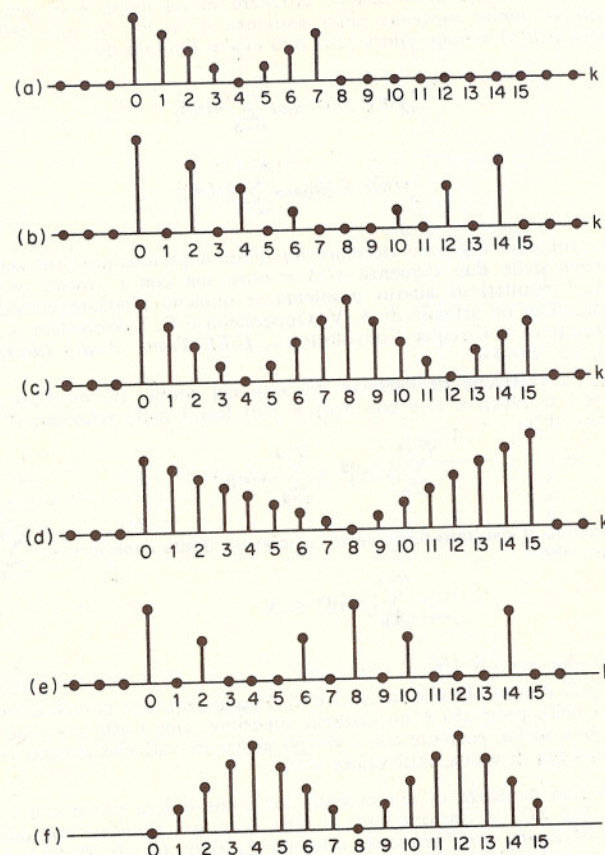


Fig. P3.18-2

19. Uno dei modi di realizzare una convoluzione circolare discreta di due sequenze di durata finita consiste nel moltiplicare le loro DFT e calcolare la DFT inversa del risultato. In particolare, indicando con $X(k)$, $Y(k)$ e $H(k)$ le DFT su N punti delle sequenze lunghe N $x(n)$, $y(n)$ e $h(n)$, e se è

$$Y(k) = X(k)H(k) \quad (\text{P3.19-1})$$

risulta

$$y(n) = \sum_{m=0}^{N-1} x(m)h((n-m))_N, \quad n = 0, 1, \dots, N-1 \quad (\text{P3.19-2})$$

Nel realizzare una convoluzione circolare in tal modo è in generale utile disporre di un limite superiore sulla sequenza di uscita. Se una delle due sequenze $x(n)$ o $h(n)$ è nota, allora $y(n)$ può essere limitata da

$$|y(n)| \leq |x|_{\max} \sum_{m=0}^{N-1} |h(m)| \quad (\text{P3.19-3})$$

o

$$|y(n)| \leq |h|_{\max} \sum_{m=0}^{N-1} |x(m)| \quad (\text{P3.19-4})$$

In questo problema vogliamo ricercare un limite superiore per $y(n)$ senza conoscere nessuna delle due sequenze $x(n)$ e $h(n)$, ma con i vincoli $|x(n)| \leq 1$ e $|H(k)| \leq 1$. I risultati di questo problema, e qualche ulteriore considerazione, sono presentati in un articolo di A. V. Oppenheim e C. J. Weinstein, « A Bound on the Output of a Circular Convolution », *IEEE Trans. Audio Electroacoust.*, June 1969, pp. 344-348.

- (a) Usando la relazione di Parseval derivata nel probl. 16, mostrare che con $|x(n)| \leq 1$ e $|H(k)| \leq 1$, e con $x(n)$ e $y(n)$ legati dalle relazioni (P3.19-3) e (P3.19-4), si ha

$$\sum_{n=0}^{N-1} |y(n)|^2 \leq \sum_{n=0}^{N-1} |x(n)|^2$$

- (b) Combinando il risultato della parte (a) con un limite superiore su $\sum_{n=0}^{N-1} |x(n)|^2$, mostrare che

$$\sum_{n=0}^{N-1} |y(n)|^2 \leq N$$

e perciò che $|y(n)| \leq \sqrt{N}$.

Si può dimostrare che, se $x(n)$ e $y(n)$ sono sequenze complesse, il limite derivato nella parte (b) è un estremo superiore, cioè esiste per ciascuna sequenza una scelta, coerente con i vincoli assegnati, tale che almeno un punto nella sequenza di uscita ha il valore \sqrt{N} .

20. Si consideri una sequenza di durata finita $x(n)$, che è zero per $n < 0$ e $n \geq N$, con N pari. Sia $X(z)$ la trasformata z di $x(n)$. Sono qui riportate due tabelle. In tab. P3.20-1 si hanno sette sequenze ottenute da $x(n)$. In tab. P3.20-2 si hanno nove sequenze ottenute da $X(z)$. Per ogni sequenza in tab. P3.20-1 si trovi la sua DFT in tab. P3.20-2. La dimensione della trasformata considerata deve essere maggiore o uguale della lunghezza della sequenza $g_k(n)$. Solo a titolo di esempio si assuma che $x(n)$ possa essere rappresentata dall'involuppo mostrato in fig. P3.20.

21. Sia $f(t)$ una funzione a tempo continuo reale, limitata in banda, periodica. Il periodo di $f(t)$ è P , così che $f(t) = f(t + sP)$ per ogni s intero. Gli unici termini non nulli nella rappresentazione in serie di Fourier complessa di $f(t)$ corrispondono a frequenze tra $-2\pi M/P$ e $2\pi M/P$, cioè

$$f(t) = \sum_{r=-M}^M a_r e^{j(2\pi r t/P)}$$

Inoltre, a_M è reale.

Si genera una sequenza $x_1(n)$ campionando $f(t)$ con periodo di campionamento T_1 , dove

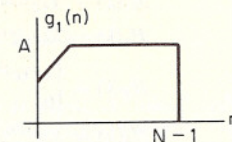
$$x_1(n) = f(nT_1)$$

e

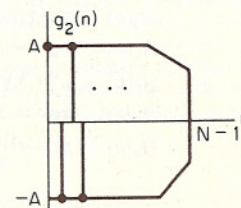
$$T_1 = \frac{P}{2M}$$

Tab. P3.20-1

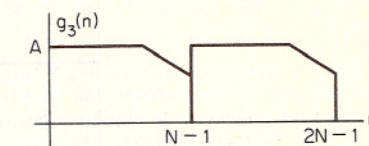
$$g_1(n) = x(N-1-n)$$



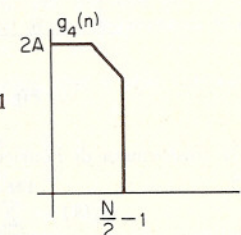
$$g_2(n) = (-1)^n x(n)$$



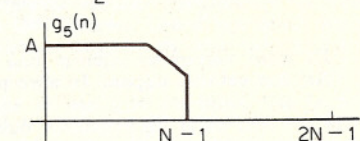
$$g_3(n) = \begin{cases} x(n), & 0 \leq n \leq N-1 \\ x(n-N), & N \leq n \leq 2N-1 \\ 0, & \text{altrove} \end{cases}$$



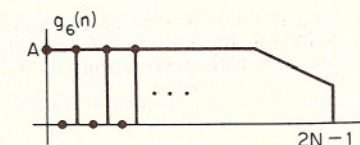
$$g_4(n) = \begin{cases} x(n) + x(n+N/2), & 0 \leq n \leq N/2-1 \\ 0, & \text{altrove} \end{cases}$$



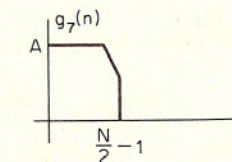
$$g_5(n) = \begin{cases} x(n), & 0 \leq n \leq N-1 \\ 0, & N \leq n \leq 2N-1 \\ 0, & \text{altrove} \end{cases}$$



$$g_6(n) = \begin{cases} x\left(\frac{n}{2}\right), & n \text{ pari} \\ 0, & n \text{ dispari} \end{cases}$$



$$g_7(n) = x(2n)$$



Tab. P3.20-2

$$\begin{aligned}
 H_1(k) &= X(e^{j2\pi k/N}) \\
 H_2(k) &= X(e^{j2\pi k/2N}) \\
 H_3(k) &= \begin{cases} 2X(e^{j2\pi k/2N}), & k \text{ pari} \\ 0, & k \text{ dispari} \end{cases} \\
 H_4(k) &= X(e^{j2\pi k/(2N-1)}) \\
 H_5(k) &= 0.5[X(e^{j2\pi k/N}) + X(e^{j2\pi(k+N/2)/N})] \\
 H_6(k) &= X(e^{j4\pi k/N}) \\
 H_7(k) &= e^{j2\pi k/N} X(e^{-j2\pi k/N}) \\
 H_8(k) &= X(e^{j2\pi/N}(k+N/2)) \\
 H_9(k) &= X(e^{-j2\pi k/N})
 \end{aligned}$$

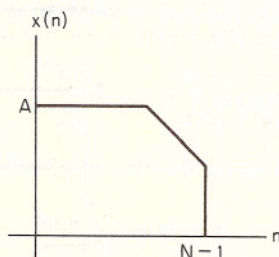


Fig. P3.20

Sia $X_1(k)$ la trasformata di Fourier discreta di un periodo di $x_1(n)$ partendo da $n = 0$, cioè

$$X_1(k) = \sum_{n=0}^{2M-1} x_1(n) e^{-j2\pi nk/2M}$$

Da $x_1(n)$ vorremmo ottenere una sequenza $x_2(n)$ che corrisponda a campionare $f(t)$ con velocità doppia. In altre parole,

$$x_2(n) = f(nT_2)$$

dove $T_2 = T_1/2 = P/4M$. Sia $X_2(k)$ la trasformata di Fourier discreta di un periodo, partendo da $n = 0$, della sequenza periodica $x_2(n)$. Determinare come si può ottenere $X_2(k)$ direttamente da $X_1(k)$. Giustificare chiaramente la risposta.

22. Sia $x(n)$ una sequenza di durata infinita, $X(z)$ la sua trasformata z , e $x_1(n)$ una sequenza di durata finita N la cui DFT sulla base di N punti è $X_1(k)$. Determinare la relazione tra $x(n)$ e $x_1(n)$ se $X(z)$ e $X_1(k)$ sono legate da

$$X_1(k) = X(z)|_{z=W_N^{-k}}, \quad k = 0, 1, 2, \dots, N-1$$

dove $W_N = e^{-j2\pi/N}$

23. Sia $X(e^{j\omega})$ la trasformata di Fourier della sequenza $x(n) = (1/2)^n u(n)$. Sia $y(n)$ una sequenza di durata finita lunga 10, cioè $y(n) = 0$ per $n < 0$, e $y(n) = 0$ per $n \geq 10$. La DFT su 10 punti di $y(n)$, indicata con $Y(k)$, corrisponde a 10 campioni equispaziati di $X(e^{j\omega})$, cioè $Y(k) = X(e^{j2\pi k/10})$. Determinare $y(n)$.

24. Siano $x(n)$ e $y(n)$ l'ingresso e l'uscita di un sistema IIR stabile e causale con equazione alle differenze della forma

$$y(n) = \sum_{k=1}^p a_k y(n-k) + x(n) \quad (\text{P3.24-1})$$

Vogliamo determinare, usando una DFT su N punti, N valori della risposta in frequenza del sistema $H(e^{j\omega})$ equispaziati sul circolo unitario, cioè per $\omega_k = (2\pi/N)k$, $k = 0, 1, \dots, N-1$. Una possibilità consiste nel generare la risposta all'impulso ed applicare i risultati del probl. 23. Per la classe di sistemi caratterizzati dalla relazione (P3.24-1), esiste però un metodo più semplice.

- (a) Assumendo che sia $p < N$, mostrare come si possono calcolare gli N valori richiesti di $H(e^{j\omega})$ dai coefficienti dell'equazione alle differenze (P3.24-1) usando una sola DFT su N punti e qualche semplice calcolo aritmetico.
(b) È possibile generalizzare questo risultato alla classe di sistemi caratterizzati da equazioni alle differenze della forma

$$y(n) = \sum_{k=1}^p a_k y(n-k) + \sum_{k=0}^q b_k x(n-k)?$$

25. Si consideri una sequenza $x(n)$ di durata finita N tale che $x(n) = 0$ per $n < 0$ e per $n > N-1$. Vogliamo calcolare i valori della sua trasformata z , $X(z)$, in M punti equispaziati sul circolo unitario. Uno dei campioni deve cadere in $z = 1$. Il numero di valori M è minore della durata della sequenza N , cioè $M < N$. Determinare e giustificare un procedimento per ottenere gli M valori di $X(z)$ calcolando una sola volta la DFT su M punti di una sequenza di M punti ricavata da $x(n)$.

26. Si considerino due sequenze di durata finita $x(n)$ e $y(n)$, entrambe nulle per $n < 0$ e con

$$\begin{aligned}
 x(n) &= 0, & n &\geq 8 \\
 y(n) &= 0, & n &\geq 20
 \end{aligned}$$

Le DFT su 20 punti di ciascuna di tali sequenze vengono moltiplicate tra loro e si calcola poi la DFT inversa. Sia $r(n)$ la DFT inversa. Specificare quali punti in $r(n)$ corrispondono a punti che si sarebbero ottenuti da una convoluzione lineare di $x(n)$ e $y(n)$.

27. Vogliamo filtrare una sequenza molto lunga di dati con un filtro FIR la cui risposta all'impulso ha durata 50. Desideriamo realizzare tale filtro con una FFT usando la tecnica di sovrapposizione ed estrazione. Per far ciò: (1) le sezioni di ingresso devono essere sovrapposte di V campioni, e (2) dall'uscita corrispondente a ciascuna sezione dobbiamo estrarre M campioni in modo tale che, quando tali campioni di ogni sezione sono giunti assieme, la sequenza risultante sia l'uscita del filtro desiderata. Si assuma che i segmenti di ingresso siano costituiti da 100 campioni e che la dimensione della DFT sia di 128 ($= 2^7$) punti. Si assuma inoltre che gli indici della sequenza di uscita dalla convoluzione circolare vadano da 0 a 127.

- (a) Determinare V .
(b) Determinare M .
(c) Determinare l'indice dell'inizio e della fine degli M punti estratti; cioè, determinare quali dei 128 punti della convoluzione circolare vanno estratti per essere aggiunti al risultato della sezione precedente.

28. È stata proposta (J. L. Vernet, « Real Signals Fast Fourier Transform: Storage Capacity and Step Number Reduction by Means of an Odd Discrete Fourier Transform », *Proc. IEEE*, Oct. 1971, pp. 1531-1532) una trasformata di Fourier

discreta modificata (MDFT) che calcola campioni della trasformata z sul circolo unitario in posizione diversa da quelli calcolati con la DFT. In particolare, se $X_M(k)$ indica la MDFT di $x(n)$,

$$X_M(k) = X(z)|_{z=e^{j2\pi k/(N+M/2)}}, \quad k = 0, 1, 2, \dots, N-1$$

Si assuma N pari.

- (a) La MDFT su N punti di una sequenza $x(n)$ corrisponde alla DFT su N punti di una sequenza $x_M(n)$ che si costruisce facilmente a partire da $x(n)$. Determinare $x_M(n)$ in termini di $x(n)$.
- (b) Se $x(n)$ è reale, i punti della DFT non sono tutti indipendenti fra loro, poiché la DFT è coniugata simmetrica, cioè $X(k) = X^*((-k))_N$. Analogamente, se $x(n)$ è reale, i punti della MDFT non sono tutti indipendenti fra loro. Determinare, per $x(n)$ reale, la relazione tra i punti di $X_M(k)$.

- (c) (1) Sia $R(k) = X_M(2k)$, cioè $R(k)$ contiene i punti con indice pari di $X_M(k)$. Dalla risposta alla parte (b), mostrare che $X_M(k)$ può essere ricostruita a partire da $R(k)$.

- (2) $R(k)$ può essere considerata come la MDFT su $N/2$ punti di una sequenza $r(n)$ di $N/2$ punti. Determinare una semplice espressione che legghi $r(n)$ direttamente a $x(n)$.

Secondo quanto riportato nelle parti (b) e (c), la MDFT su N punti di una sequenza reale $x(n)$ può essere calcolata formando $r(n)$ da $x(n)$ e quindi calcolando la MDFT su $N/2$ punti di $r(n)$. Le due parti seguenti hanno lo scopo di mostrare che la MDFT può essere usata per realizzare una convoluzione lineare.

- (d) Si considerino tre sequenze, $x_1(n)$, $x_2(n)$ e $x_3(n)$, tutte di durata N . Siano $X_{1M}(k)$, $X_{2M}(k)$ e $X_{3M}(k)$, rispettivamente, le MDFT delle tre sequenze. Se

$$X_{3M}(k) = X_{1M}(k)X_{2M}(k)$$

esprimere $x_3(n)$ in termini di $x_1(n)$ e $x_2(n)$. L'espressione deve avere la forma di una singola sommatoria su una « combinazione » di $x_1(n)$ e $x_2(n)$, allo stesso modo di (ma non identica a) una convoluzione circolare.

- (e) Si può chiamare il risultato della parte (d) una convoluzione circolare modificata. Se le sequenze $x_1(n)$ e $x_2(n)$ sono entrambe nulle per $n \geq N/2$, mostrare che la convoluzione circolare modificata di $x_1(n)$ e $x_2(n)$ è identica alla convoluzione lineare di $x_1(n)$ e $x_2(n)$.

29. Vogliamo realizzare un filtro numerico passa-basso sezionando l'ingresso, calcolando la DFT di ogni sezione, moltiplicando per la DFT della risposta all'impulso del filtro, calcolando la DFT inversa, e mettendo assieme le sezioni. Il numero di valori non nulli nella risposta all'impulso è M e la durata di una sezione di ingresso è $N + M - 1$.

Sono stati proposti due metodi per ottenere la DFT $H(k)$, su $(N + M - 1)$ punti, che rappresenta il filtro. In entrambi i metodi si inizia considerando una DFT su M punti $H_M(k)$, data da

$$H_M(k) = \begin{cases} 1, & 0 \leq k < M/4 \\ 1, & 3M/4 < k \leq M-1 \\ 0, & \text{altrove} \end{cases}$$

Si assuma che M sia divisibile per 4. La DFT inversa su M punti di $H_M(k)$ è indicata con $h_M(n)$.

Metodo A: $H(k)$ è la DFT su $(N + M - 1)$ punti di $h_A(n)$ definita da

$$h_A(n) = \begin{cases} h_M(n), & 0 \leq n \leq M-1 \\ 0, & M-1 < n \leq N+M-2 \end{cases}$$

Metodo B: $H(k)$ è la DFT su $(N + M - 1)$ punti di $h_B(n)$ definita da

$$h_B(n) = \begin{cases} h_M(n), & 0 \leq n \leq M/2 - 1 \quad (\text{si assuma } M \text{ pari}) \\ h_M(n - N + 1), & N + M/2 - 1 \leq n \leq N + M - 2 \\ 0, & \text{altrove} \end{cases}$$

- (a) Disegnare $h_A(n)$ e $h_B(n)$. L'ingresso è sezionato sovrapponendo $M - 1$ punti come è mostrato in fig. P3.29.

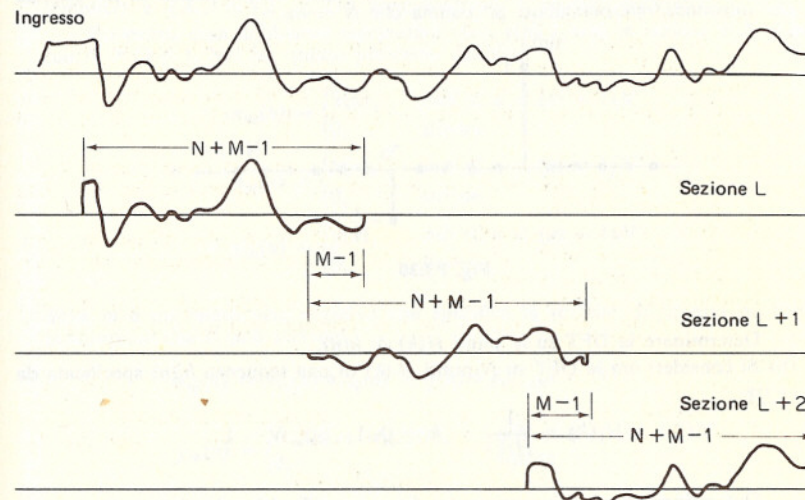


Fig. P3.29

Per collegare nuovamente insieme le sezioni filtrate, si possono usare due metodi.

Metodo 1: Le sezioni filtrate sono messe assieme giuntando solo gli ultimi N punti.

Metodo 2: Le sezioni filtrate sono messe assieme eliminando gli ultimi $M/2$ e i primi $(M/2) - 1$ punti e giuntando i rimanenti N punti.

- (b) Con ciascuno dei due metodi per riunire le sezioni si hanno due possibilità per ottenere $H(k)$. Siano A-1, A-2, B-1 e B-2 le quattro possibili scelte. Daranno tutte e quattro come risultato un filtro lineare tempo-invariante? Giustificare la risposta.
- (c) Per ciascuna scelta che dà come risultato un filtro lineare tempo-invariante, determinare e disegnare la risposta all'impulso. Quale di questi è il « miglior » filtro passa-basso?

30. Sorge spesso il problema in cui un segnale $x(n)$ è stato filtrato da un sistema lineare tempo-invariante che dà come risultato un segnale distorto $y(n)$ e si desidera ricostruire il segnale originale. Ciò può essere spesso fatto elaborando $y(n)$ con un sistema lineare tempo-invariante la cui risposta all'impulso è tale che la risposta all'impulso complessiva dei due sistemi in cascata sia un campione unitario. Questo procedimento è generalmente detto *filtraggio inverso*.

Abbiamo visto nel par. 3.8 il procedimento per realizzare un filtro FIR usando la DFT. Il procedimento comporta, tra l'altro, la moltiplicazione della DFT dell'ingresso, $X(k)$ (o di sezioni dell'ingresso), per $H(k)$, la DFT della risposta all'impulso del sistema, per ottenere $Y(k)$, la DFT dell'uscita.

È diffusa l'idea, sbagliata, che la risposta all'impulso del filtro inverso sia la sequenza $h_i(n)$, la cui DFT è $1/H(k)$. Scopo di questo problema è far vedere perché tale idea è errata.

Si consideri un sistema lineare tempo-invariante con risposta all'impulso $h(n)$ data da

$$h(n) = \delta(n) - \frac{1}{2}\delta(n - n_0)$$

come mostrato in fig. P3.30. Questo sistema è un esempio ideale di un sistema che introduce riverberazione. Si assuma che $N = 4n_0$.

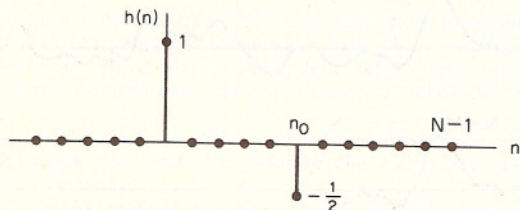


Fig. P3.30

- (a) Determinare la DFT su N punti $H(k)$ di $h(n)$.
- (b) Si consideri ora la DFT su N punti $H_1(k)$ di una sequenza $h_1(n)$ specificata da

$$H_1(k) = \frac{1}{H(k)} \quad k = 0, 1, \dots, N-1$$

Determinare $h_1(n)$. [Suggerimento: Se si ha difficoltà nel valutare direttamente la sommatoria della IDFT, esprimere $H_1(k)$ come polinomio in W_N^{nk} e osservare che gli $h_1(n)$ sono i coefficienti di W_N^{nk} .]

- (c) Disegnare $h_1(n)$ determinato in (b).
 - (d) Valutando la convoluzione lineare di $h(n)$ e $h_1(n)$, mostrare che $h(n) * h_1(n)$ non è un campione unitario $\delta(n)$ e, di conseguenza, $h_1(n)$ non è la risposta all'impulso del sistema inverso.
 - (e) Calcolare e disegnare la convoluzione circolare su N punti di $h(n)$ e $h_1(n)$.
 - (f) Determinare la risposta all'impulso $h_i(n)$ del sistema inverso per $h(n)$. Ciò può essere fatto in vari modi. Uno è notare che, se $H(z)$ e $H_i(z)$ indicano le trasformate z di $h(n)$ e $h_i(n)$, è $H_i(z) = 1/H(z)$. La trasformata z inversa di $H_i(z)$ può poi essere calcolata con una divisione lunga.
 - (g) Facendo riferimento al probl. 23 determinare e verificare numericamente la relazione tra $h_i(n)$ e $h(n)$.
31. Abbiamo visto che i filtri FIR possono essere realizzati usando la DFT. Abbiamo anche visto che i filtri IIR possono essere realizzati ricorsivamente. In questo problema considereremo la realizzazione di filtri IIR mediante la DFT.
- In particolare, assumiamo che $x(n)$ sia la sequenza di ingresso ad un sistema lineare, invariante alla traslazione, stabile e causale, caratterizzato dalla risposta all'impulso $h(n)$. Assumiamo che sia $x(n) = 0$ per $n < 0$. Indicheremo con $y(n)$ la sequenza di uscita (fig. P3.31).

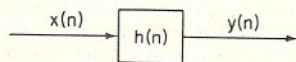


Fig. P3.31

Assumeremo inoltre che $H(z)$, la trasformata z di $h(n)$, abbia solo poli. Perciò, possiamo scrivere

$$H(z) = \frac{1}{D(z)} = \frac{1}{1 + \sum_{i=1}^Q d_i z^{-i}} = \sum_{i=1}^Q \frac{A_i}{1 - z_i z^{-1}} \quad (\text{P3.31-1})$$

dove $\{z_i, i = 1, 2, \dots, Q\}$ rappresentano i poli del sistema (per ipotesi tutti poli semplici) e $\{A_i, i = 1, 2, \dots, Q\}$ rappresentano i residui associati.

Per prima cosa dobbiamo segmentare $x(n)$, $y(n)$ e $h(n)$ in sezioni di N punti, con $N \geq Q + 1$. Nel far questo possiamo definire

$$\begin{aligned} x_m(n) &= \begin{cases} x(n), & mN \leq n < (m+1)N \\ 0, & \text{altrove} \end{cases} \\ h_m(n) &= \begin{cases} h(n), & mN \leq n < (m+1)N \\ 0, & \text{altrove} \end{cases} \\ y_m(n) &= \begin{cases} y(n), & mN \leq n < (m+1)N \\ 0, & \text{altrove} \end{cases} \end{aligned}$$

dove m è un intero non negativo che specifica la sezione. In questo problema mostreremo che si può scrivere

$$w_m(n) = x_m(n) * h_0(n) + g(n) * y_{m-1}(n - N) \quad (\text{P3.31-2})$$

$$y_m(n) = \begin{cases} w_m(n) + w_{m-1}(n), & mN \leq n < (m+1)N \\ 0, & \text{altrove} \end{cases}$$

dove $g(n)$ è una sequenza di durata $\leq N$.

- (a) Assumendo che una tale $g(n)$ possa essere trovata, illustrare le operazioni che devono essere eseguite per calcolare $y_m(n)$ nella (P3.31-2), usando tecniche basate sulla DFT. È sufficiente un semplice diagramma di flusso.

Per verificare la (P3.31-2), supporremo che esista una $g(n)$ e quindi mostreremo che essa (1) ha durata finita e (2) può essere valutata in base a parametri noti.

- (b) Dimostrare che

$$H(z) = \frac{Y(z)}{X(z)} = \frac{H_0(z)}{1 - z^{-N}G(z)} = \frac{1}{D(z)}$$

dove $H_0(z)$ è la trasformata z di $h_0(n)$, la prima sezione della risposta all'impulso.

- (c) Calcolando esplicitamente $H_0(z)$ e usando la relazione ricavata in (b), trovare $G(z)$ in termini di $D(z)$, $\{z_i\}$, $\{A_i\}$, N e Q . Verificare che essa è la trasformata z di una sequenza di durata Q [Nota: Non è necessario calcolare $g(n)$].
- (d) Usando i risultati di questo problema, mostrare che nel caso generale in cui $H(z) = C(z)/D(z)$, con $C(z)$ e $D(z)$ polinomi in z^{-1} , la rete può essere realizzata usando la DFT.

Commento: Le quattro parti di questo problema possono essere tutte svolte indipendentemente. Il significato del risultato ricavato in questo problema è che un filtro IIR può essere realizzato per mezzo di convoluzioni circolari e della DFT. Se usiamo come termine di paragone il numero di moltiplicazioni richieste per realizzare il sistema discreto, allora $D(z)$ dovrebbe essere un polinomio di grado 95 (circa), prima che questa tecnica diventi preferibile alla realizzazione diretta dell'equazione alle differenze che lega $y(n)$ a $x(n)$. Perciò questo risultato è di interesse più teorico che pratico.

Questo problema è basato su un articolo di B. Gold e K. L. Jordan, « A Note on Digital Filter Synthesis », *Proc. IEEE (Letters)*, Vol. 56, Oct. 1968, pp. 1717-1718.

32. Le espressioni (3.53) e (3.54) rappresentano due modi alternativi di calcolare una DFT bidimensionale usando una DFT monodimensionale. La (3.53) corrisponde a trasformare prima ogni colonna di $x(m, n)$ e poi ogni riga del risultato. La (3.54) corrisponde a trasformare prima le righe e poi le colonne. Per approfondire questi due procedimenti alternativi è istruttivo considerare un semplice esempio.
- (a) Calcolare la DFT bidimensionale della sequenza $\delta(m, n - 1)$ usando la (3.53).
 (b) Ripetere la parte (a) usando la (3.54).
33. Mostrare che la DFT bidimensionale di una sequenza di area finita corrisponde a campioni della trasformata z bidimensionale. Specificare in particolare la posizione di questi campioni nello spazio (z_1, z_2) .
34. Nel probl. 16 è stata verificata la relazione di Parseval per la DFT monodimensionale. Determinare e verificare la corrispondente relazione per la DFT bidimensionale.
35. Esiste un insieme di proprietà di simmetria per la DFT bidimensionale che è simile a quello del caso monodimensionale ricavato nel par. 3.6.3. Queste proprietà sono ancora basate sulla scomposizione di una sequenza $x(m, n)$ nelle sue componenti coniugata simmetrica e coniugata antisimmetrica. Specificamente, ricalcando le (3.31)–(3.35), definiamo

$$x_{ep}(m, n) = \frac{1}{2}\{x[((m))_M, ((n))_N] + x^*[((-m))_M, ((-n))_N]\} \mathcal{R}_{M,N}(m, n)$$

$$x_{op}(m, n) = \frac{1}{2}\{x[((m))_M, ((n))_N] - x^*[((-m))_M, ((-n))_N]\} \mathcal{R}_{M,N}(m, n)$$

- (a) Mostrare che valgono le seguenti proprietà per $x(m, n)$ e la sua DFT $X(k, l)$:

Sequenza	DFT
1. $x^*(m, n)$	$X^*[((-k))_M, ((-l))_N] \mathcal{R}_{M,N}(k, l)$
2. $x^*[(((-m))_M, ((-n))_N)] \mathcal{R}_{M,N}(m, n)$	$X^*(k, l)$
3. $x_{ep}(m, n)$	$\text{Re}[X(k, l)]$
4. $x_{op}(m, n)$	$j \text{Im}[X(k, l)]$
5. $\text{Re}[x(m, n)]$	$X_{ep}(k, l)$
6. $j \text{Im}[x(m, n)]$	$X_{op}(k, l)$

- (b) Mostrare che se $x(m, n)$ è reale, allora

$$(1) \text{Re}[X(k, l)] = \text{Re}\{X[((-k))_M, ((-l))_N] \mathcal{R}_{M,N}(k, l)\}.$$

$$(2) |X(k, l)| = |X[(((-k))_M, ((-l))_N)] \mathcal{R}_{M,N}(k, l)|$$

$$(3) \text{Im}[X(k, l)] = -\text{Im}\{X[(((-k))_M, ((-l))_N)] \mathcal{R}_{M,N}(k, l)\}.$$

4. USO DEI GRAFI DI FLUSSO E MATRICI PER LA RAPPRESENTAZIONE DEI FILTRI NUMERICI

4.0 INTRODUZIONE

Nei capitoli 1 e 2 abbiamo discusso la rappresentazione dei sistemi discreti lineari invarianti alla traslazione sia in termini di equazioni alle differenze, le quali mettono in relazione le sequenze di ingresso e uscita di quei sistemi, che in termini di funzioni di trasferimento che mettono invece in relazione le trasformate z delle medesime sequenze. In quei capitoli abbiamo dunque trattato soltanto la relazione ingresso-uscita dei sistemi. Se tuttavia si vuole realizzare un filtro numerico o su calcolatore o mediante circuiti « ad hoc » (cioè in « hardware »), occorre esprimere la relazione di ingresso-uscita con un algoritmo di calcolo, cioè mediante un insieme di operazioni o blocchi elementari. La scelta più conveniente per questi elementi di base dell'algoritmo è, nel nostro caso, quella delle tre operazioni fondamentali di addizione, ritardo e moltiplicazione per una costante. L'algoritmo di calcolo per realizzare il filtro è pertanto definito da una struttura o rete consistente in una interconnessione di queste tre operazioni di base. A scopo illustrativo si consideri un sistema con una funzione caratteristica della forma:

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{1 - \sum_{k=1}^N a_k z^{-k}} = \frac{Y(z)}{X(z)} \quad (4.1)$$

L'equazione alle differenze che lega ingresso e uscita si scrive in modo facile e diretto a partire dalla funzione caratteristica ed è data da:

$$y(n) = \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^M b_k x(n-k) \quad (4.2)$$

La (4.2) è interpretabile direttamente come un algoritmo di calcolo nel quale i valori ritardati dell'ingresso sono moltiplicati per i coefficienti b_k , i valori ritardati dell'uscita sono moltiplicati per i coefficienti a_k e tutti

i prodotti risultanti vengono poi sommati. In alternativa, come vedremo in questo capitolo, esiste una infinita varietà di strutture che danno luogo alla stessa relazione fra i valori di ingresso $x(n)$ e i valori in uscita $y(n)$.

Nei prossimi paragrafi descriveremo le strutture dei filtri numerici in termini di diagrammi a blocchi, grafi di flusso e matrici. Inoltre prenderemo in considerazione più di una struttura di base. Infatti strutture di base che sono equivalenti (per quanto riguarda le relazioni ingresso-uscita) se i coefficienti e le variabili hanno precisione infinita, possono avere caratteristiche molto diverse quando la precisione è limitata. In questo capitolo discuteremo gli effetti della rappresentazione con precisione finita dei coefficienti di un filtro. Gli effetti del troncamento e arrotondamento dei calcoli intermedi saranno invece discussi nel cap. 9.

Le strutture che noi useremo per le reti di elaborazione numerica* rappresentano, in sostanza, il flusso dei segnali nella realizzazione di un filtro numerico. A questo scopo abbiamo a disposizione una teoria ormai ben collaudata che è quella dei grafi lineari di flusso di segnale. Nell'ambito di questa teoria esistono numerose ed utili proprietà che riguardano le reti suddette ed una di queste, il teorema di Tellegen, sarà presentata e discussa nel par. 4.6.

4.1 RAPPRESENTAZIONE DELLE RETI NUMERICHE MEDIANTE GRAFI DI FLUSSO DI SEGNALE

La realizzazione di un filtro numerico richiede che siano disponibili valori passati dell'uscita, dell'ingresso ed eventualmente di sequenze intermedie. Ciò implica la necessità di ritardare o memorizzare questi valori passati. Inoltre occorrono dispositivi per moltiplicare i campioni ritardati per i coefficienti, e dispositivi per sommare fra loro i prodotti risultanti. Il filtro può pertanto essere realizzato o usando i registri di memoria e le unità aritmetica e di controllo di un calcolatore « general-purpose », o progettando speciali circuiti che eseguano i calcoli richiesti. Nel primo caso la struttura del filtro può essere pensata come la specificazione di un algoritmo di calcolo, da cui si deriva un programma di calcolo (soluzione « software »). Nell'altro caso è spesso conveniente pensare alla struttura del filtro come qualcosa che specifica una particolare configurazione circuitale (soluzione « hardware »).

Corrispondentemente alle operazioni base richieste per realizzare un filtro numerico, gli elementi base necessari per rappresentare graficamente

* Nel seguito verranno usate con lo stesso significato sia l'espressione « rete di elaborazione numerica » che l'espressione « rete numerica » (n.d.t.).

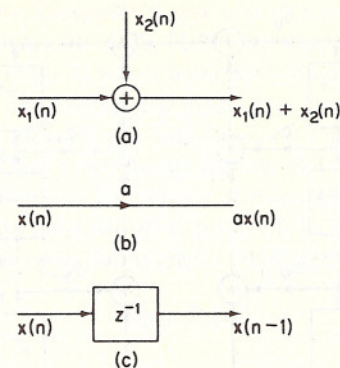


Fig. 4.1 Simboli di diagrammi a blocchi per una rete di elaborazione numerica: (a) somma di due sequenze; (b) moltiplicazione di una sequenza per una costante; (c) ritardo unitario.

un'equazione alle differenze sono un sommatore, un ritardatore ed un moltiplicatore per una costante. Simboli comunemente usati sono quelli mostrati in fig. 4.1. Specificamente la fig. 4.1 (a) rappresenta un dispositivo per sommare una all'altra due sequenze, la fig. 4.1 (b) rappresenta un dispositivo per moltiplicare una sequenza per una costante, e la fig. 4.1 (c) rappresenta un dispositivo per memorizzare il valore precedente di una sequenza. La rappresentazione usata per l'elemento ritardatore di un singolo campione deriva dal fatto che la trasformata z di $x(n-1)$ è semplicemente z^{-1} volte la trasformata z di $x(n)$.

Come esempio di rappresentazione di un'equazione alle differenze per mezzo di questi elementi, si consideri l'equazione del secondo ordine

$$y(n) = a_1 y(n-1) + a_2 y(n-2) + b x(n)$$

La rete corrispondente a questa equazione è mostrata nella fig. 4.2. In termini di un programma per un calcolatore la fig. 4.2 mostra esplicitamente

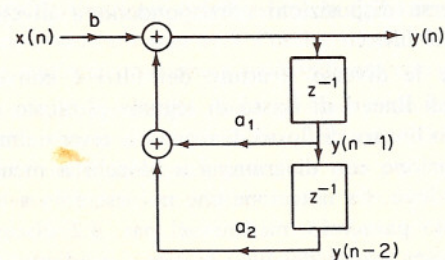


Fig. 4.2 Esempio di una rete di elaborazione numerica.

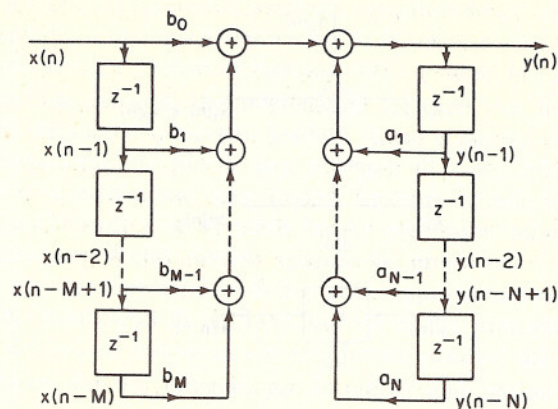


Fig. 4.3 Rappresentazione con diagramma a blocchi per la più generale equazione alle differenze di ordine N .

che si devono memorizzare le variabili $y(n-1)$ ed $y(n-2)$ ed anche le costanti a_1 , a_2 e b . Inoltre si vede che un dato campione in uscita si calcola formando i prodotti $a_1 y(n-1)$ e $a_2 y(n-2)$, sommandoli insieme e poi sommando il risultato al prodotto $b x(n)$. In termini di realizzazione circuitale « ad hoc » (« hardware »), la fig. 4.2 indica che occorre provvedere a dispositivi di memorizzazione di variabili e costanti, nonché a dispositivi di moltiplicazione e addizione. Pertanto diagrammi come quelli della fig. 4.2 servono a raffigurare sia la complessità di un algoritmo di filtraggio numerico sia la quantità di « hardware » richiesta per realizzare il filtro.

Come ulteriore esempio mostriamo in fig. 4.3 una rappresentazione con diagramma a blocchi della più generale equazione alle differenze (4.2). La rete di fig. 4.3 è una esplicita rappresentazione grafica di questa equazione alle differenze. Tuttavia essa può essere modificata o ricomposta in molti modi diversi senza che cambi la funzione di trasferimento complessiva. Queste diverse disposizioni corrispondono a diverse strutture per la realizzazione del filtro.

Nel discutere le diverse strutture dei filtri è conveniente usare la simbologia dei grafi lineari di flusso di segnale piuttosto che i diagrammi a blocchi. Un grafo lineare di flusso di segnale è essenzialmente equivalente a una rappresentazione con diagramma a blocchi a meno di pochissime differenze di notazione. La notazione che noi useremo sarà presentata nel prosieguo di questo paragrafo, mentre nel par. 4.2 discuteremo l'equivalenza tra le rappresentazioni dei filtri numerici mediante grafi di flusso e matrici.

Un grafo di flusso di segnale è una rete di rami orientati che si connettono in corrispondenza di nodi [1,2]. Ad ogni nodo è associata una variabile o valore del nodo. Il valore associato col nodo k è w_k . Il ramo (jk) indica un ramo che ha origine nel nodo j e termina nel nodo k , con la direzione da j a k indicata con una freccia sul ramo stesso. Ciò è mostrato nella fig. 4.4. Ogni ramo ha un segnale di ingresso e un segnale di uscita. Il segnale di ingresso dal nodo j al ramo (jk) è il valore del nodo w_j e il segnale di uscita dal ramo jk al nodo k è indicato con v_{jk} . La dipendenza dell'uscita di un ramo dall'ingresso è indicata con

$$v_{jk} = f_{jk}[w_j] \quad (4.3)$$

dove $f_{jk} []$ è l'operatore che trasforma l'ingresso nell'uscita di un ramo.

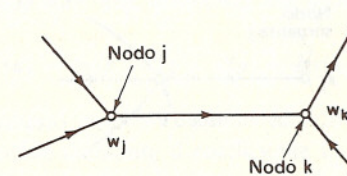


Fig. 4.4 Esempio di nodi e rami in un grafo di flusso di segnale.

Per rappresentare l'immissione nel grafo di ingressi o sorgenti esterne usiamo i *nodi sorgente*. Un nodo sorgente non ha rami entranti. È generalmente conveniente numerare i nodi sorgente separatamente dai nodi di rete. Il valore di nodo al nodo sorgente j sarà indicato con x_j e l'uscita di un ramo che connette il nodo sorgente j al nodo di rete k sarà indicata con s_{jk} . Un esempio di nodo sorgente è raffigurato in fig. 4.5.

Come gli ingressi in un grafo possono essere rappresentati mediante nodi sorgente, così è spesso conveniente rappresentare le uscite da un grafo mediante i *nodi destinazione*, cioè nodi che hanno soltanto rami entranti. Un nodo destinazione è raffigurato in fig. 4.6. Il valore nel nodo destinazione k sarà indicato con y_k e l'uscita di un ramo che connette il nodo di rete j al nodo destinazione k sarà indicato con r_{jk} .

Per definizione, il valore del nodo ad ogni nodo è dato dalla somma delle uscite di tutti i rami che entrano in quel nodo. Talora è conveniente, dal punto di vista delle notazioni, assumere che per ogni coppia di nodi di rete esistano rami in ciascuna direzione e che ogni nodo sorgente sia connesso ad ogni nodo di rete, benché chiaramente alcune delle uscite dei rami possano essere nulle. Con questa notazione e assumendo che si abbiano N nodi di rete, numerati da 1 ad N , M nodi sorgente numerati da

1 ad M , e P nodi destinazione numerati da 1 a P , l'insieme delle equazioni rappresentate dal grafo è

$$w_k = \sum_{j=1}^N v_{jk} + \sum_{j=1}^M s_{jk}, \quad k = 1, 2, \dots, N \quad (4.4a)$$

(nodi di rete) (nodi sorgente)

$$y_k = \sum_{j=1}^N r_{jk} \quad k = 1, 2, \dots, P \quad (4.4b)$$

(nodi di rete)

Per vedere un esempio di come questi concetti di grafi di flusso possono essere applicati alla rappresentazione delle equazioni alle differenze, consideriamo il diagramma a blocchi del filtro numerico del primo ordine di fig. 4.7 (a). Un grafo di flusso di segnale corrispondente a questa rete è

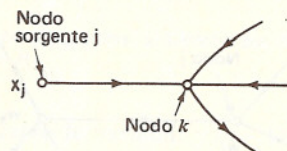


Fig. 4.5 Rappresentazione di un nodo sorgente.

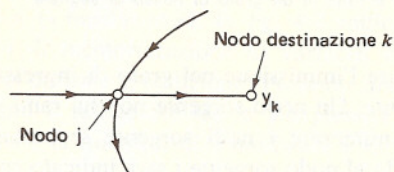


Fig. 4.6 Rappresentazione di un nodo destinazione.

mostrato in fig. 4.7 (b). In questo caso le variabili di ramo sono sequenze. C'è un unico nodo sorgente connesso al nodo 1 ed un unico nodo destinazione connesso al nodo 3. Scrivendo le equazioni (4.4) per questo grafo otteniamo

$$w_1(n) = s_{11}(n) + v_{41}(n)$$

$$w_2(n) = v_{12}(n)$$

$$w_3(n) = v_{23}(n) + v_{43}(n)$$

$$w_4(n) = v_{24}(n)$$

$$y(n) = w_3(n)$$

Dalla fig. 4.7 (b) notiamo che le uscite dei rami sono

$$s_{11}(n) = x(n)$$

$$v_{12}(n) = f_{12}(w_1) = w_1(n)$$

$$v_{23}(n) = f_{23}(w_2) = w_2(n)$$

$$v_{43}(n) = f_{43}(w_4) = bw_4(n)$$

$$v_{41}(n) = f_{41}(w_4) = aw_4(n)$$

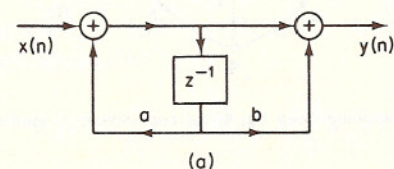
$$v_{24}(n) = f_{24}(w_2) = w_2(n-1) \quad (\text{ritardo})$$

$$y(n) = w_3(n)$$

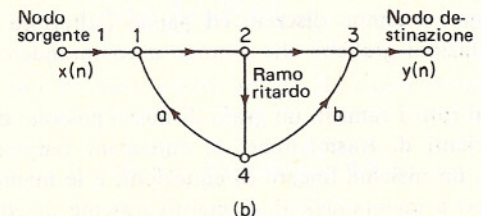
Queste equazioni possono essere risolte per $y(n)$ in termini di $x(n)$, dando luogo alla singola equazione alle differenze del primo ordine

$$y(n) = cy(n-1) + x(n) + bx(n-1)$$

Osserviamo che in questo caso tutti i rami eccetto uno (il ramo 2,4) possono essere rappresentati mediante il coefficiente di trasmissione del ramo; ovvero il segnale di uscita è semplicemente l'ingresso moltiplicato per una costante. Il ramo (2,4) è invece rappresentato da un operatore di ritardo. In effetti, in generale, $f_{jk} []$ sta a indicare un operatore che trasforma una sequenza di ingresso in un ramo in una sequenza di uscita. Nel caso dei sistemi lineari tempo-invarianti e a tempo discreto caratterizzati da equazioni alle differenze, il grafo di flusso di segnale può anche



(a)



(b)

Fig. 4.7 (a) Rappresentazione con diagramma a blocchi di un filtro numerico del primo ordine; (b) struttura del grafo di flusso di segnale corrispondente al diagramma a blocchi in (a).

rappresentare relazioni fra trasformate z . In questo caso ogni ramo è caratterizzabile con la sua funzione di trasferimento, cioè con un coefficiente di trasmissione che è funzione di z . Pertanto

$$V_{jk}(z) = F_{jk}(z)W_j(z) \quad (4.5)$$

Per tali grafi il coefficiente di trasmissione di ogni ramo verrà scritto in prossimità della freccia che indica la direzione del ramo. Per convenienza si assumerà che un ramo senza esplicita indicazione del coefficiente di trasmissione abbia tale coefficiente pari ad uno. Inoltre è talora conveniente indicare le variabili di nodo come sequenze invece che come trasformate z , nel qual caso è inteso che coefficienti di trasmissione pari a z^{-1} implicano un ritardo unitario della sequenza di ingresso. Il grafo dell'esempio precedente è mostrato ancora nella fig. 4.8 nella forma che sarà usata d'ora in poi.

Il confronto della fig. 4.7(a) con la fig. 4.8 mostra che esiste una diretta corrispondenza fra i rami nella rete (diagramma a blocchi) e i rami nel grafo di flusso. Infatti, la sola differenza importante tra i due casi è che i nodi nel grafo di flusso corrispondono nella rete sia a punti di semplice connessione che a sommatore. Per esempio, i nodi 1 e 3 corrispondono a sommatore e i nodi 2 e 4 corrispondono a punti di connessione nella rete originale. I grafi di flusso di segnale servono a fornire una raffi-

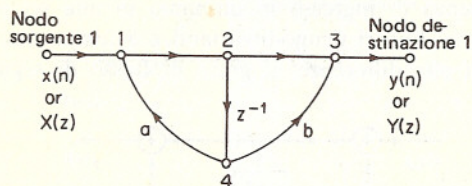


Fig. 4.8 Grafo di flusso di segnale della fig. 4.7(b) con indicati i coefficienti di trasmissione dei rami.

gurazione dei sistemi a tempo discreto ed hanno l'ulteriore vantaggio di consentire manipolazioni grafiche che aiutano a comprendere il funzionamento della rete.

Nei casi in cui tutti i rami in un grafo di flusso possono essere rappresentati con coefficienti di trasmissione, le equazioni rappresentative del grafo costituiscono un insieme lineare di equazioni, e le manipolazioni del grafo corrispondono a manipolazioni di questo insieme di equazioni. Pertanto una descrizione alternativa di un grafo lineare di flusso è proprio in termini di questo insieme di equazioni, corrispondente a una rappresentazione matriciale, come vedremo nel prossimo paragrafo.

4.2 RAPPRESENTAZIONE MATRICIALE DELLE RETI NUMERICHE

Le equazioni (4.4) rappresentate con un grafo di flusso di segnale ed espresse in termini di trasformate z diventano

$$W_k(z) = \sum_{j=1}^N V_{jk}(z) + \sum_{j=1}^M S_{jk}(z), \quad k = 1, 2, \dots, N \quad (4.6a)$$

$$Y_k(z) = \sum_{j=1}^N R_{jk}(z) \quad k = 1, 2, \dots, P \quad (4.6b)$$

Nel caso in cui il grafo rappresenta un sistema lineare invariante, alla traslazione ogni ramo può essere rappresentato mediante un coefficiente di trasmissione. Sarà conveniente, in generale, assumere che i rami dai nodi sorgente ai nodi di rete e da questi ai nodi destinazione abbiano un coefficiente di trasmissione costante, indipendente cioè da z . Questa ipotesi non comporta perdita di generalità in quanto, se necessario, si può sempre inserire, in connessione diretta con un nodo sorgente, un nodo di rete che abbia rami, verso altri nodi, con coefficienti di trasmissione non costanti. Un procedimento simile può essere usato, se necessario, per i nodi destinazione. Pertanto

$$V_{jk}(z) = F_{jk}(z)W_j(z) \quad (4.7)$$

e

$$S_{jk}(z) = b_{jk}X_j(z) \quad (4.8)$$

$$R_{jk}(z) = c_{jk}W_j(z) \quad (4.9)$$

Sostituendo le espressioni (4.7) ÷ (4.9) nelle (4.6) otteniamo l'insieme di equazioni lineari algebriche

$$W_k(z) = \sum_{j=1}^N F_{jk}(z)W_j(z) + \sum_{j=1}^M b_{jk}X_j(z) \quad (4.10a)$$

$$Y_k(z) = \sum_{j=1}^N c_{jk}W_j(z) \quad (4.10b)$$

Queste equazioni possono essere scritte in modo compatto nella forma matriciale come

$$\mathbf{W}(z) = \mathbf{F}^t(z)\mathbf{W}(z) + \mathbf{B}^t\mathbf{X}(z) \quad (4.11a)$$

$$\mathbf{Y}(z) = \mathbf{C}^t\mathbf{W}(z) \quad (4.11b)$$

dove $\mathbf{W}(z)$ è un vettore colonna di valori $W_k(z)$, $k = 1, 2, \dots, N$; $\mathbf{X}(z)$ è un vettore colonna di valori $X_j(z)$, $j = 1, 2, \dots, M$, e $\mathbf{Y}(z)$ è un vettore colonna

di valori $Y_j(z)$, $j = 1, 2, \dots, P$. La matrice $\mathbf{F}'(z)$ è la trasposta della matrice $N \times N \mathbf{F}(z)$ data da

$$\mathbf{F}(z) = \{F_{kj}(z)\} \quad (4.12)$$

Per rami che non esistano nel grafo di flusso, o che, equivalentemente, abbiano coefficiente di trasmissione zero, il corrispondente valore $F_{kj}(z)$ nella matrice è zero. \mathbf{B}' è la trasposta $N \times M$ della matrice

$$\mathbf{B} = \{b_{kj}\} \quad (4.13)$$

e \mathbf{C}' è la trasposta $P \times N$ della matrice $\mathbf{C} = \{c_{kj}\}$. La trasposizione delle matrici, indicata con t , si rende necessaria nelle eq. (4.11) per ragioni di compatibilità tra le convenzioni adottate per gli indici nei grafi di flusso e nelle matrici.

L'eq. (4.11 a) può essere risolta rispetto a $\mathbf{W}(z)$ con una inversione di matrice dando luogo a

$$\begin{aligned} \mathbf{W}(z) &= [\mathbf{I} - \mathbf{F}'(z)]^{-1} \mathbf{B}' \mathbf{X}(z) \\ &= \mathbf{T}'(z) \mathbf{X}(z) \end{aligned} \quad (4.14a)$$

dove

$$\mathbf{T}'(z) = [\mathbf{I} - \mathbf{F}'(z)]^{-1} \mathbf{B}' = \{T_{jk}(z)\} \quad (4.14b)$$

$\mathbf{T}(z)$ è chiamata la *matrice funzione di trasferimento* del sistema. Come conseguenza della (4.14) il segnale al k -mo nodo, $W_k(z)$, si può esprimere come

$$W_k(z) = \sum_{j=1}^M T_{jk}(z) X_j(z) \quad (4.15)$$

cioè ogni variabile di nodo si può esprimere come combinazione lineare dei valori dei nodi sorgente. Se soltanto un nodo sorgente (il nodo sorgente a) è diverso da zero, con valore $X_a(z)$, e se esiste un unico nodo destinazione con valore $Y(z)$, tale che $Y(z) = \mathbf{C}' \mathbf{W}(z)$, allora l'uscita $Y(z)$ è data da

$$Y(z) = \mathbf{C}' \mathbf{W}(z) = \mathbf{C}' \mathbf{T}' X_a(z) \quad (4.16)$$

per cui il sistema è caratterizzato dalla funzione di trasferimento

$$H(z) = \mathbf{C}' \mathbf{T}' \quad (4.17)$$

Nel caso in cui la funzione caratteristica di ciascun ramo è al più del primo ordine, vale a dire è una costante moltiplicativa oppure una costante moltiplicativa associata a un ritardo unitario, allora gli elementi della matrice $\mathbf{F}'(z)$ nell'eq. (4.11 a) sono o una costante o una costante per z^{-1} . È conveniente separare gli elementi della matrice che non sono associati a un ritardo da quelli che lo sono, così che $\mathbf{F}'(z)$ può esprimersi come

$$\mathbf{F}'(z) = \mathbf{F}'_c + z^{-1} \mathbf{F}'_d \quad (4.18)$$

dove \mathbf{F}'_c ed \mathbf{F}'_d sono matrici $N \times N$. L'equazione matriciale (4.11 a) si può quindi scrivere

$$\mathbf{W}(z) = \mathbf{F}'_c \mathbf{W}(z) + z^{-1} \mathbf{F}'_d \mathbf{W}(z) + \mathbf{B}' \mathbf{X}(z) \quad (4.19)$$

Analogamente l'espressione di $\mathbf{T}'(z)$ diventa

$$\mathbf{T}'(z) = [\mathbf{I} - \mathbf{F}'_c - z^{-1} \mathbf{F}'_d]^{-1} \mathbf{B}' \quad (4.20)$$

dove \mathbf{I} è la matrice identità. Poiché \mathbf{F}'_c e \mathbf{F}'_d sono costanti, ovvero indipendenti da z , la trasformata z inversa della (4.19) è data da

$$\mathbf{w}(n) = \mathbf{F}'_c \mathbf{w}(n) + \mathbf{F}'_d \mathbf{w}(n-1) + \mathbf{B}' \mathbf{x}(n) \quad (4.21a)$$

Inoltre, la (4.11b) implica che

$$\mathbf{y}(n) = \mathbf{C}' \mathbf{w}(n) \quad (4.21b)$$

Ovviamente le espressioni (4.21) possono essere scritte direttamente a partire dal grafo di flusso, o, viceversa, è possibile disegnare il grafo di flusso direttamente a partire da tali espressioni.

ESEMPIO. A titolo di esempio, si consideri il sistema del primo ordine di fig. 4.9. L'insieme di equazioni rappresentato implicitamente da tale grafo è

$$\begin{bmatrix} w_1(n) \\ w_2(n) \\ w_3(n) \\ w_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & a_1 \\ 1 & 0 & 0 & 0 \\ 0 & b_0 & 0 & b_1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1(n) \\ w_2(n) \\ w_3(n) \\ w_4(n) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1(n-1) \\ w_2(n-1) \\ w_3(n-1) \\ w_4(n-1) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} [x(n)]$$

$$y(n) = [0 \quad 0 \quad 1 \quad 0] \begin{bmatrix} w_1(n) \\ w_2(n) \\ w_3(n) \\ w_4(n) \end{bmatrix} = w_3(n) \quad (4.22)$$

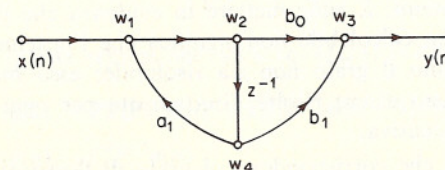


Fig. 4.9 Grafo di flusso di segnale per un sistema del primo ordine; le equazioni corrispondenti sono date dalle (4.22).

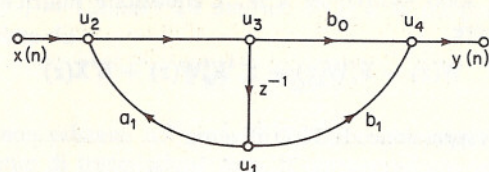


Fig. 4.10 Grafo di flusso di fig. 4.9 con diversa numerazione dei nodi.

Chiaramente la forma delle matrici F_c' e F_d' dipende dal modo con cui si ordinano le equazioni, oppure, in maniera equivalente, dalla numerazione dei nodi utilizzata. In fig. 4.10 abbiamo rappresentato il grafo di fig. 4.9 con le variabili di nodo (indicate con u_k) numerate diversamente. Per questo grafo si ottiene l'insieme di equazioni

$$\begin{bmatrix} u_1(n) \\ u_2(n) \\ u_3(n) \\ u_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ a_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ b_1 & 0 & b_0 & 0 \end{bmatrix} \begin{bmatrix} u_1(n) \\ u_2(n) \\ u_3(n) \\ u_4(n) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(n-1) \\ u_2(n-1) \\ u_3(n-1) \\ u_4(n-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} [x(n)]$$

$$y(n) = [0 \quad 0 \quad 0 \quad 1] \begin{bmatrix} u_1(n) \\ u_2(n) \\ u_3(n) \\ u_4(n) \end{bmatrix} = u_4(n) \quad (4.23)$$

Dall'esame del grafo di flusso di fig. 4.9 o delle eq. (4.22), che sono del tutto equivalenti, si vede chiaramente che le variabili di nodo non possono essere generate in sequenza, cioè prima w_1 , poi w_2 , etc.; infatti, ad esempio, w_4 deve essere noto per poter calcolare w_1 . D'altra parte, lo stesso diagramma di flusso con la diversa numerazione dei nodi di fig. 4.10 può essere calcolato con le variabili di nodo generate in sequenza.

In alcuni casi non è possibile riordinare i nodi di un grafo in modo che le variabili di nodo possano essere generate in sequenza. Un grafo di flusso di questo tipo è detto *non calcolabile*. Un semplice esempio di grafo non calcolabile è rappresentato in fig. 4.11, in cui tutti i coefficienti di trasmissione sono costanti. È bene mettere in evidenza che il fatto che il grafo di flusso sia non calcolabile *non* significa che l'insieme delle equazioni che rappresentano il grafo non sia risolvibile; esso indica soltanto che queste non possono essere risolte direttamente per ogni variabile di nodo in maniera consecutiva.

Nelle eq. (4.23), che corrispondono al grafo di flusso di fig. 4.10, si può notare che la matrice F_c' ha solo zeri sopra la diagonale principale e che inoltre tutti gli elementi della diagonale principale sono nulli; ciò

non si verifica nelle (4.22), che corrispondono al grafo di flusso di fig. 4.9. Come si vedrà nel probl. 3 di questo capitolo, condizione necessaria e sufficiente per la calcolabilità di un grafo di flusso è che i nodi possano

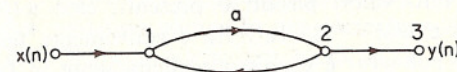


Fig. 4.11 Esempio di grafo non calcolabile.

essere numerati in modo che la matrice F_c' abbia tutti zeri al di sopra della diagonale principale e che gli elementi di questa siano tutti nulli. Si può anche dimostrare [3] che una condizione necessaria e sufficiente per la calcolabilità di un grafo di flusso, equivalente alla precedente, è che non ci siano maglie del grafo prive di rami con ritardo. Nel grafo di fig. 4.11, ad esempio, c'è una maglia priva di rami con ritardo, e di conseguenza il grafo non è calcolabile.

L'insieme delle (4.23) può essere ottenuto dalle (4.22) permutando le variabili di nodo; in notazione matriciale ciò può essere ottenuto per mezzo di una trasformazione lineare del vettore $w(n)$ nel vettore $u(n)$, cioè

$$u(n) = Pw(n) \quad (4.24)$$

dove P è una matrice costante $N \times N$ non singolare. Nell'esempio precedente P vale

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (4.25)$$

La permutazione delle variabili è un esempio semplice di un principio più generale che può essere messo in evidenza con la rappresentazione matriciale. In generale, se P è una qualsiasi matrice non singolare, si può scrivere

$$w(n) = P^{-1}u(n) \quad (4.26)$$

e sostituendo questa espressione nelle eq. (4.21) si ottiene

$$\begin{aligned} u(n) &= PF_c'P^{-1}u(n) + PF_c'P^{-1}u(n-1) + PB'x(n) \\ y(n) &= C'P^{-1}u(n) \end{aligned} \quad (4.27)$$

Le eq. (4.27) hanno la stessa forma delle (4.21), ma corrispondono ad un diverso grafo di flusso o rappresentazione di rete; perciò è possibile

una varietà di realizzazioni di rete per una stessa funzione di trasferimento. Questo fatto importante sarà fondamentale nei par. 4.3 e 4.4.

Nelle eq. (4.21) i valori attuali delle variabili di nodo sono espressi in funzione dei loro valori passati e presenti, cioè $w(n)$ è espresso in funzione di $w(n)$ e $w(n-1)$. Talvolta è conveniente trattare rappresentazioni di rete in cui $w(n)$ è espresso esplicitamente in funzione solo dei valori precedenti delle variabili di nodo e dei valori correnti dell'ingresso. Ciò corrisponde al caso in cui la matrice F'_c di (4.21) è uguale a zero. Con le equazioni di rete espresse in questa forma si può calcolare il valore del vettore di nodo $w(n_1)$ in ogni istante n_1 dato il valore del vettore $w(n_0)$ ad un certo istante n_0 e il vettore di ingresso $s(n)$ per $n_0 \leq n \leq n_1$ (si veda il probl. 5 di questo capitolo). Questa rappresentazione è simile a quella comunemente detta rappresentazione con le variabili di stato, sebbene nella nostra formulazione il numero delle variabili di stato (nodi) sia generalmente più grande di quello strettamente necessario per rappresentare la rete. Una rappresentazione con variabili di stato di questo tipo può essere ottenuta da una rappresentazione nella forma delle eq. (4.21). In particolare, la (4.21 a) può essere scritta come

$$[I - F'_c]w(n) = F'_d w(n-1) + B'x(n)$$

Assumendo che la matrice $[I - F'_c]$ sia non singolare, l'equazione precedente si può risolvere rispetto a $w(n)$, ottenendo

$$w(n) = [I - F'_c]^{-1} F'_d w(n-1) + [I - F'_c]^{-1} B'x(n)$$

Se si pone, per definizione,

$$D = [I - F'_c]^{-1} B' \quad (4.28a)$$

e

$$A = [I - F'_c]^{-1} F'_d \quad (4.28b)$$

si ottiene la rappresentazione matriciale

$$w(n) = Aw(n-1) + Dx(n) \quad (4.29a)$$

$$y(n) = C'w(n) \quad (4.29b)$$

Si può dimostrare (v. probl. 4 di questo capitolo) che, se il sistema è calcolabile, allora la matrice $[I - F'_c]$ è non singolare. Di conseguenza si può trovare una rappresentazione matriciale della forma (4.29) per ogni sistema calcolabile.

La trasformazione delle variabili di nodo secondo la (4.26) corrisponde ad una trasformazione del grafo di flusso; vale a dire si modifica

la struttura della rete mantenendo invariate le relazioni di ingresso-uscita. Siccome ci sono molte trasformazioni del tipo (4.26), ci possono essere molte realizzazioni di rete per una stessa funzione di trasferimento. In pratica c'è un certo numero di configurazioni di rete che sono le più usate comunemente. Nel prossimo paragrafo considereremo alcune delle forme più comuni di reti. Sebbene queste reti siano legate l'una all'altra attraverso trasformazioni lineari del tipo della (4.26), ci sembra più utile giungere a queste forme standard di reti per altra via; per fare questo è utile considerare separatamente i sistemi IIR e quelli FIR.

4.3 LE STRUTTURE DI RETE FONDAMENTALI PER SISTEMI IIR

Nei due paragrafi precedenti, in cui si è parlato della rappresentazione dei sistemi lineari a tempo discreto invarianti alla traslazione sotto forma di reti, è stato messo in chiaro che ad ogni funzione di trasferimento razionale corrisponde una molteplicità di configurazioni di rete. Una prima considerazione circa la scelta tra queste differenti realizzazioni è la complessità di calcolo: vale a dire che le reti con il minimo numero di moltiplicatori per una costante e con il minimo numero di rami di ritardo sono spesso le preferite, in quanto la moltiplicazione è un'operazione che richiede tempo e ogni elemento di ritardo corrisponde all'occupazione di un registro di memoria. Di conseguenza, una riduzione nel numero dei moltiplicatori significa un aumento della velocità, ed una diminuzione nel numero dei ritardi significa una riduzione nell'occupazione di memoria. D'altra parte, gli effetti della lunghezza finita dei registri nell'effettiva realizzazione « hardware » di filtri numerici dipendono dalla struttura, e talvolta è preferibile usare una struttura che non abbia un numero minimo di moltiplicatori e ritardi, ma che sia meno sensibile agli effetti della lunghezza finita dei registri. Per questi motivi è importante analizzare alcune delle strutture più comunemente usate; in questo paragrafo analizzeremo i sistemi IIR e nel par. 4.5 si parlerà dei sistemi FIR.

4.3.1 Forma diretta

Si consideri una funzione di trasferimento razionale della forma

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (4.30)$$

e si ricordi che l'ingresso e l'uscita di tale sistema soddisfano l'equazione alle differenze finite

$$y(n) = \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^M b_k x(n-k) \quad (4.31)$$

Siccome questa equazione alle differenze può essere scritta direttamente in base all'espressione della funzione di trasferimento nella forma (4.30), la rete corrispondente alla (4.31) è detta la realizzazione in *forma diretta I* del sistema caratterizzato dalla (4.30). Questa rete è rappresentata in fig. 4.12. Disegnando la rete, abbiamo assunto per comodità $M=N$. Chiamamente, se ciò non avviene, i coefficienti di trasmissione di alcuni rami saranno nulli. Bisogna osservare che in fig. 4.12 abbiamo disegnato il grafo della rete in modo che ogni nodo abbia al più due ingressi. Questa convenzione dà luogo a un numero di nodi maggiore del necessario, ma è coerente col fatto che nelle realizzazioni numeriche, sia « hardware » che « software », la somma di più di due numeri è ottenuta attraverso le somme dei numeri a due a due fatte separatamente.

Poiché l'insieme dei coefficienti b_k corrisponde al polinomio al numeratore e l'insieme dei coefficienti a_k corrisponde al polinomio al denominatore di $H(z)$, possiamo interpretare il filtro di fig. 4.12 come l'insieme di due reti poste in cascata, la prima delle quali realizza gli zeri e la seconda i poli. Nel caso di sistemi lineari invarianti alla traslazione, la relazione ingresso-uscita complessiva della cascata è indipendente dall'ordine in cui sono disposti i sottosistemi. Questa proprietà suggerisce un

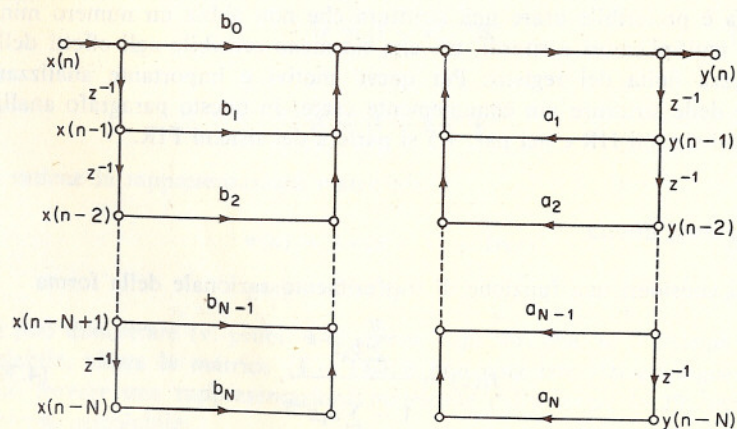


Fig. 4.12 Realizzazione in forma diretta I di una equazione alle differenze finite di ordine N .

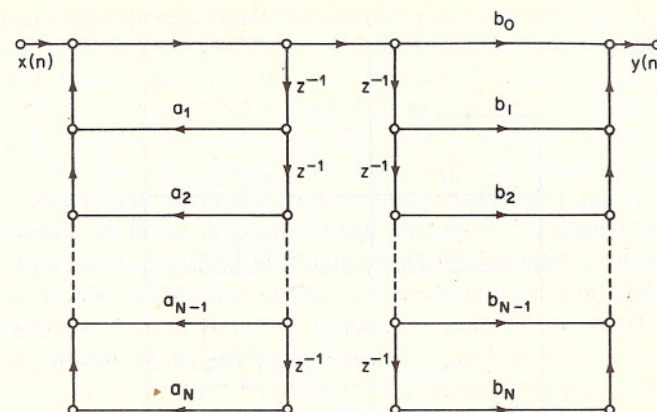


Fig. 4.13 La rete di fig. 4.12, dove l'ordine in cui sono realizzati poli e zeri è invertito.

secondo tipo di realizzazione in forma diretta: in particolare, se si realizzano prima i poli di $H(z)$, corrispondenti alla parte destra di fig. 4.12, e poi gli zeri, si ottiene la rete di fig. 4.13.

Osserviamo che le due linee di rami con coefficiente di trasmissione z^{-1} hanno lo stesso ingresso, per cui solo una linea è necessaria; di conseguenza, la rete di fig. 4.13 può essere ridisegnata come in fig. 4.14. Questa configurazione viene spesso indicata come *forma diretta II*. Notiamo che essa ha il minimo numero di rami (M o N , il maggiore dei due) con coefficiente di trasmissione z^{-1} , cioè per questa realizzazione della funzione di

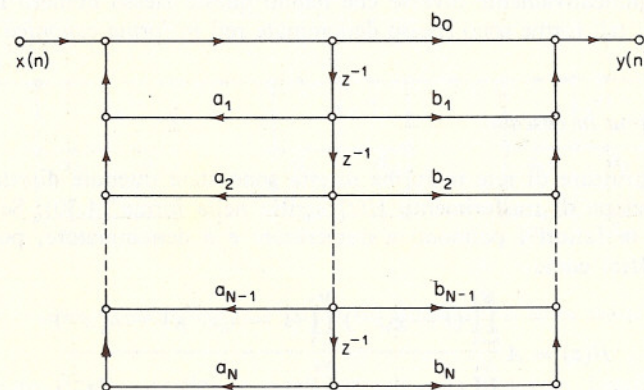


Fig. 4.14 La rete di fig. 4.13 in cui le due linee di ritardi sono fuse in un'unica linea. Il tipo di rete che ne risulta è indicata come *forma diretta II* ed ha il minimo numero possibile di ritardi.

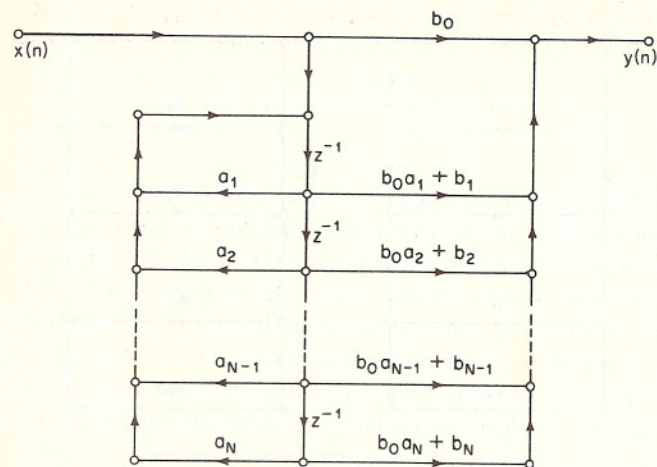


Fig. 4.15 Esempio di struttura diversa da quella di fig. 4.14, ma che pure ha il numero minimo di ritardi.

trasferimento $H(z)$ espressa dalla (4.30) è richiesto il numero minimo di registri di ritardo.

Si può ottenere un altro esempio di struttura avente il minimo numero di ritardi scrivendo la funzione di trasferimento nella forma

$$H(z) = b_0 + \sum_{r=1}^M \frac{(b_0 a_r + b_r) z^{-r}}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (4.32)$$

Il grafo di flusso corrispondente è indicato in fig. 4.15. Ci sono molte forme significativamente diverse che hanno questo stesso numero minimo di ritardi; tali forme sono spesso denominate reti in *forma canonica*.

4.3.2 Forma in cascata

Le strutture di rete in forma diretta sono state ottenute direttamente dalla funzione di trasferimento $H(z)$ scritta nella forma (4.30). Se scomponiamo in fattori i polinomi a numeratore e a denominatore, possiamo scrivere $H(z)$ come

$$H(z) = A \frac{\prod_{k=1}^{M_1} (1 - g_k z^{-1}) \prod_{k=1}^{M_2} (1 - h_k z^{-1})(1 - h_k^* z^{-1})}{\prod_{k=1}^{N_1} (1 - c_k z^{-1}) \prod_{k=1}^{N_2} (1 - d_k z^{-1})(1 - d_k^* z^{-1})} \quad (4.33)$$

dove $M = M_1 + M_2$ e $N = N_1 + N_2$. In questa espressione i fattori del primo ordine rappresentano zeri reali in g_k e poli reali in c_k , e i fattori del

secondo ordine rappresentano zeri complessi coniugati in h_k e h_k^* e poli complessi coniugati in d_k e d_k^* . Ciò rappresenta la distribuzione di poli e zeri più generale nel caso in cui tutti i coefficienti a_k e b_k della (4.30) siano reali. La formula (4.33) suggerisce un insieme di strutture consistenti nella cascata di sottosistemi del primo e secondo ordine. Chiaramente c'è una notevole libertà nella scelta della composizione dei sottosistemi e dell'ordine in cui essi sono disposti. In pratica, tuttavia, è importante costruire la forma in cascata con il minimo uso di memoria, ed inoltre le realizzazioni « hardware » sono spesso basate sull'uso ripetitivo di un'unica sezione del secondo ordine (« time sharing » o « multiplexing »). Per questi motivi è conveniente pensare in generale in termini di una forma in cascata basata sull'espressione

$$H(z) = A \prod_{k=1}^{[(N+1)/2]} \frac{1 + \beta_{1k} z^{-1} + \beta_{2k} z^{-2}}{1 - \alpha_{1k} z^{-1} - \alpha_{2k} z^{-2}} \quad (4.34)$$

dove $[(N+1)/2]$ indica il più grande numero intero contenuto in $(N+1)/2$. In questo caso abbiamo assunto che sia $M \leq N$; scrivendo $H(z)$ in questa forma si è anche assunto che i poli e zeri reali siano stati raggruppati in coppie. Se ci fosse un numero dispari di zeri reali, uno dei coefficienti β_{2k} sarebbe nullo; analogamente, se ci fosse un numero dispari di poli reali, uno dei coefficienti α_{2k} sarebbe nullo. La precedente analisi delle strutture in forma diretta rende chiaro che si possono costruire forme in cascata con minima occupazione di memoria utilizzando per ogni sottosistema del secondo ordine una realizzazione in forma diretta II. In fig. 4.16 è rappresentata una realizzazione in cascata di un sistema del sesto ordine per mezzo di sottosistemi del secondo ordine in forma diretta II.

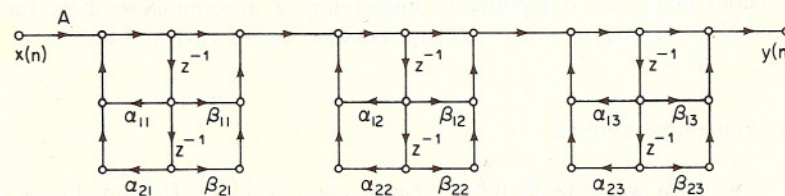


Fig. 4.16 Struttura in cascata con realizzazione dei sottosistemi in forma diretta II.

Come è stato messo in evidenza, c'è una notevole flessibilità nel modo in cui i poli e gli zeri possono essere accoppiati e nell'ordine in cui i sottosistemi del secondo ordine risultanti possono essere messi in cascata. Sebbene tali accoppiamenti ed ordinamenti siano equivalenti nel caso di

aritmetica con precisione illimitata, essi possono tuttavia differire considerevolmente in pratica, a causa degli effetti della lunghezza finita di parola.

4.3.3 Forma in parallelo

In alternativa alla scomposizione in fattori dei polinomi numeratore e denominatore, possiamo esprimere $H(z)$ con un'espansione in fratti semplici della forma

$$H(z) = \sum_{k=1}^{N_1} \frac{A_k}{1 - c_k z^{-1}} + \sum_{k=1}^{N_2} \frac{B_k(1 - e_k z^{-1})}{(1 - d_k z^{-1})(1 - d_k^* z^{-1})} + \sum_{k=0}^{M-N} C_k z^{-k} \quad (4.35)$$

Se i coefficienti a_k e b_k nella (4.30) sono reali, allora le quantità A_k , B_k , C_k , c_k ed e_k saranno tutte reali. Nel caso che sia $M < N$, il termine $\sum_{k=0}^{M-N} C_k z^{-k}$ non comparirà nella (4.35).

L'espressione (4.35) può essere interpretata come una combinazione in parallelo di sistemi del primo e secondo ordine, oppure possiamo raggruppare i poli reali a coppie in modo da poter scrivere $H(z)$ nella forma

$$H(z) = \sum_{k=0}^{M-N} C_k z^{-k} + \sum_{k=1}^{[(N+1)/2]} \frac{\gamma_{0k} + \gamma_{1k} z^{-1}}{1 - \alpha_{1k} z^{-1} - \alpha_{2k} z^{-2}} \quad (4.36)$$

Un esempio tipico di realizzazione in forma parallela con $M = N$ è riportato in fig. 4.17.

Le strutture di rete qui analizzate rappresentano le strutture fondamentali che si incontrano comunemente; ce ne sono però molte altre, che corrispondono alla molteplicità di modi in cui si possono manipolare le equazioni della rete o i polinomi numeratore e denominatore della funzione di trasferimento. Alcuni esempi possono essere visti in [4-8].

4.4 FORME TRASPOSTE

La teoria dei grafi di flusso lineari consente di trasformare in vario modo i grafi di flusso di segnale, pur lasciando invariata la relazione tra ingresso e uscita. Una di queste trasformazioni, detta *inversione del grafo di flusso* o *trasposizione*, genera un insieme di forme trasposte per i filtri. In particolare, la trasposizione di un grafo di flusso è ottenuta invertendo la direzione di tutti i rami della rete; per sistemi ad un ingresso ed un'uscita, il grafo risultante ha la stessa funzione di trasferimento del grafo originale, con l'ingresso e l'uscita scambiati. Questo teorema è una diretta conseguenza della formula del guadagno di Mason per i grafi di flusso di

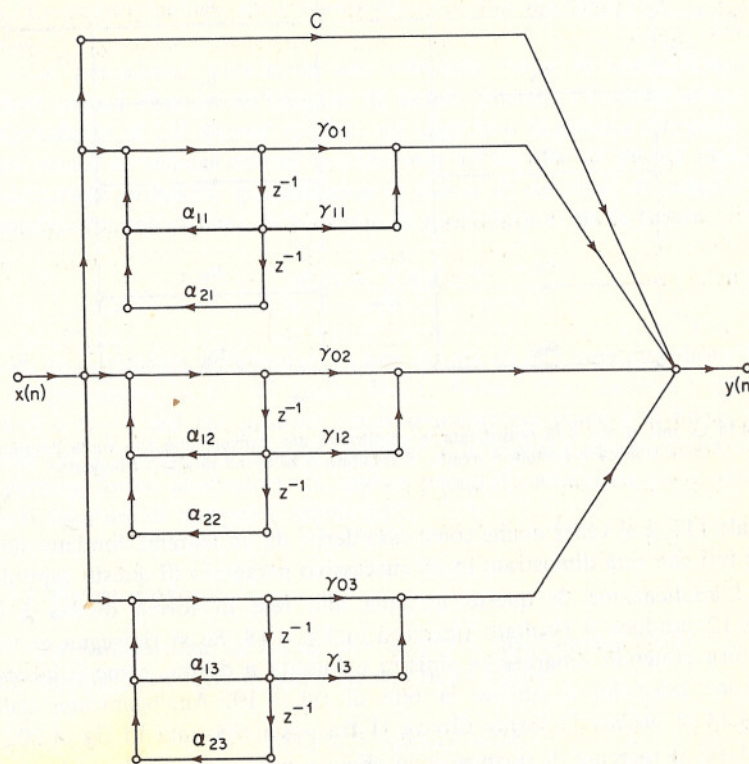


Fig. 4.17 Realizzazione in forma parallela con i poli reali e complessi riuniti in coppie.

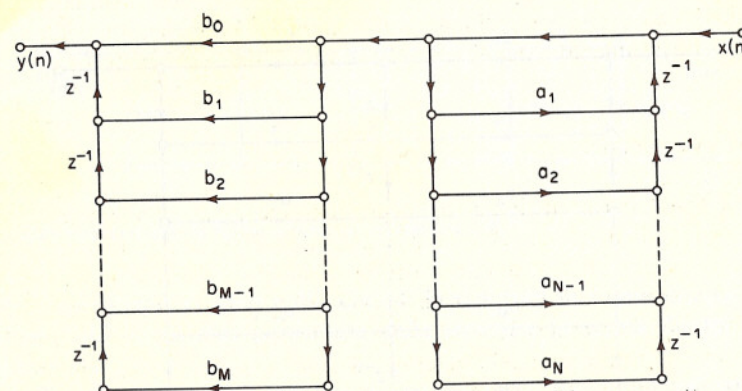


Fig. 4.18 Trasposizione della rete in forma diretta I di fig. 4.12.

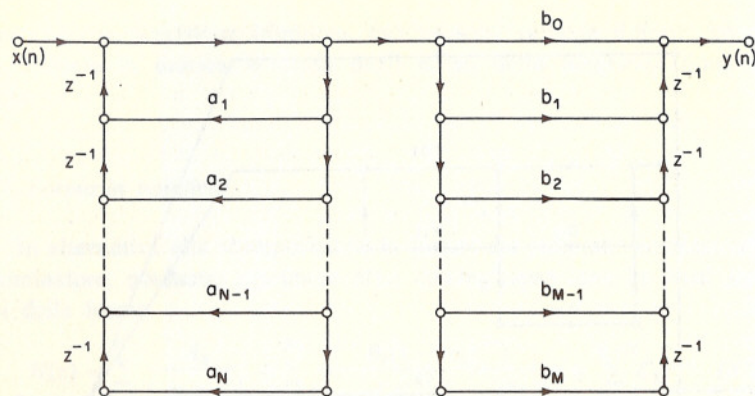


Fig. 4.19 La rete di fig. 4.18 ridisegnata in conformità alla convenzione che vuole l'ingresso sulla sinistra e l'uscita a destra. Il risultato è la forma diretta I trasposta.

segnale [1], e si vedrà anche come esso derivi da un teorema fondamentale delle reti che sarà dimostrato in un successivo paragrafo di questo capitolo.

L'applicazione di questo teorema alla rete in forma diretta I di fig. 4.12 produce il risultato riportato in fig. 4.18. Se si ridisegna questa struttura ponendo l'ingresso a sinistra e l'uscita a destra, come è la convenzione normale, si ottiene la rete di fig. 4.19. Analogamente, dalla fig. 4.14 si ottiene la forma diretta II trasposta riportata in fig. 4.20.

Dato il teorema di trasposizione appena enunciato, è chiaro che ogni configurazione di rete può essere trasposta per produrre un'altra struttura di rete, che conserva lo stesso numero di rami di ritardo e lo stesso numero di coefficienti.

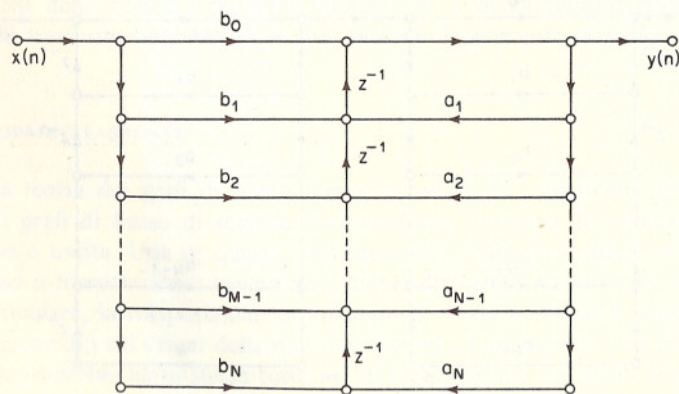


Fig. 4.20 Forma diretta II trasposta.

4.5 LE STRUTTURE DI RETE FONDAMENTALI PER SISTEMI FIR

La precedente trattazione era orientata verso la realizzazione di sistemi aventi risposta all'impulso di durata infinita. Necessariamente, la realizzazione di tali sistemi richiede un algoritmo di calcolo ricorsivo; nel caso invece di sistemi causali con risposta all'impulso di durata finita, le realizzazioni prendono generalmente la forma di algoritmi di calcolo non ricorsivi. Per tali sistemi, la funzione di trasferimento ha la forma

$$H(z) = \sum_{n=0}^{N-1} h(n)z^{-n} \quad (4.37)$$

Cioè, se la risposta all'impulso ha una durata di N campioni, allora $H(z)$ è un polinomio in z^{-1} di grado $N-1$. Perciò $H(z)$ ha $N-1$ poli in $z = 0$ e $N-1$ zeri che possono essere ovunque nel piano z al finito. Così come per i sistemi IIR, anche i sistemi FIR possono essere realizzati in numerose forme alternative; in questo paragrafo analizzeremo le più comuni strutture di rete per i sistemi FIR.

4.5.1 Forma diretta

La realizzazione in forma diretta deriva direttamente dalla relazione che esprime la somma di convoluzione nella forma

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \quad (4.38)$$

Un grafo di flusso corrispondente alla (4.38) è riportato in fig. 4.21; si può vedere che questa struttura è identica a quella di fig. 4.14 quando

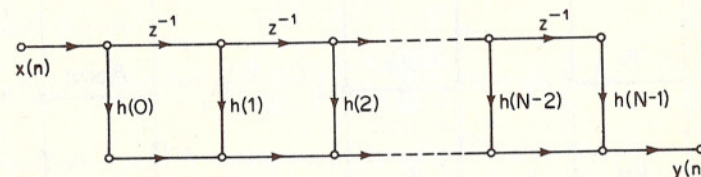


Fig. 4.21 Realizzazione in forma diretta di un sistema FIR.

tutti i coefficienti a_k sono nulli. Quindi la struttura in forma diretta per i sistemi FIR è un caso particolare della struttura in forma diretta per i sistemi IIR.

La rete di fig. 4.21 corrisponde nel modo più diretto all'ordinamento delle addizioni e moltiplicazioni richieste dalla (4.38). Chiaramente i calcoli possono essere organizzati in molte altre maniere, e quindi ci sono

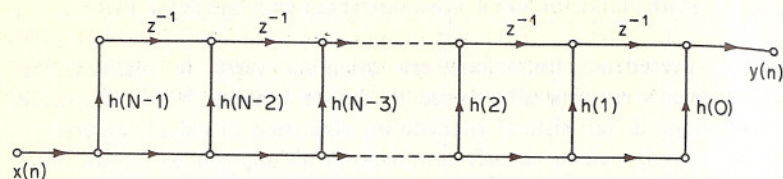


Fig. 4.22 Trasposizione della rete di fig. 4.21.

molte altre strutture di rete teoricamente equivalenti; per esempio, il teorema di trasposizione introdotto nel paragrafo precedente può essere applicato alla fig. 4.21 ottenendo la forma diretta trasposta di fig. 4.22.

4.5.2 Forma in cascata

Un'alternativa alla forma diretta si ricava scrivendo $H(z)$ sotto forma di prodotto di fattori del secondo ordine, cioè

$$H(z) = \prod_{k=1}^{[N/2]} (\beta_{0k} + \beta_{1k}z^{-1} + \beta_{2k}z^{-2}) \quad (4.39)$$

dove, se N è pari, uno dei coefficienti β_{2k} sarà zero, in relazione al fatto che se N è pari, $H(z)$ ha un numero dispari di radici reali. La rete corrispondente alla (4.39) è riportata in fig. 4.23, dove ogni fattore del secondo ordine è stato realizzato nella forma diretta illustrata in fig. 4.21.

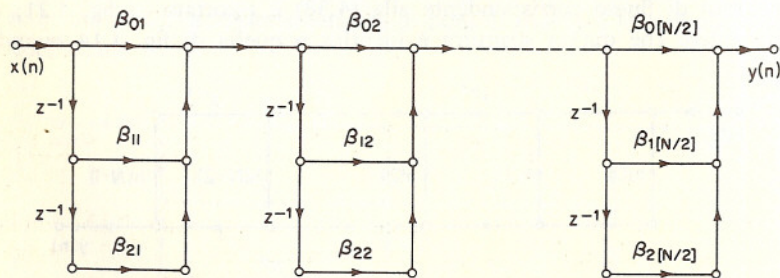


Fig. 4.23 Realizzazione di un sistema FIR nella forma in cascata.

4.5.3 Reti per sistemi FIR a fase lineare

In molte applicazioni è conveniente progettare filtri aventi fase lineare. Ad esempio, nel caso di filtri selettivi in frequenza a fase lineare, i segnali che cadono interamente all'interno della banda passante saranno

riprodotti con un ritardo uguale alla pendenza della curva di fase. Una delle più importanti caratteristiche dei sistemi FIR è che essi possono essere progettati in modo da avere fase esattamente lineare. La risposta all'impulso di un sistema FIR causale con fase lineare ha la proprietà seguente

$$h(n) = h(N-1-n) \quad (4.40)$$

Per vedere che questa condizione implica fase lineare, scriviamo la (4.37) come

$$\begin{aligned} H(z) &= \sum_{n=0}^{(N/2)-1} h(n)z^{-n} + \sum_{n=N/2}^{N-1} h(n)z^{-n} \\ &= \sum_{n=0}^{(N/2)-1} h(n)z^{-n} + \sum_{n=0}^{(N/2)-1} h(N-1-n)z^{-(N-1-n)} \end{aligned}$$

dove N è assunto pari. Per mezzo della (4.40) possiamo scrivere

$$H(z) = \sum_{n=0}^{(N/2)-1} h(n)[z^{-n} + z^{-(N-1-n)}] \quad (4.41)$$

Se N è dispari, si può facilmente dimostrare che

$$H(z) = \sum_{n=0}^{[(N-1)/2]-1} h(n)[z^{-n} + z^{-(N-1-n)}] + h\left(\frac{N-1}{2}\right)z^{-[(N-1)/2]} \quad (4.42)$$

Se calcoliamo le espressioni (4.41) e (4.42) per $z = e^{j\omega}$, otteniamo, per N pari,

$$H(e^{j\omega}) = e^{-j\omega[(N-1)/2]} \left\{ \sum_{n=0}^{(N/2)-1} 2h(n) \cos \left[\omega \left(n - \frac{N-1}{2} \right) \right] \right\}$$

e per N dispari,

$$H(e^{j\omega}) = e^{-j\omega[(N-1)/2]} \left\{ h\left(\frac{N-1}{2}\right) + \sum_{n=0}^{[(N-3)/2]} 2h(n) \cos \left[\omega \left(n - \frac{N-1}{2} \right) \right] \right\}$$

In entrambi i casi, le somme tra parentesi graffe sono reali, il che implica uno sfasamento lineare corrispondente al ritardo di $(N-1)/2$ campioni. Si osservi che per N pari $(N-1)/2$ non è un numero intero.

Le espressioni (4.41) e (4.42) implicano realizzazioni di rete in forma diretta che richiedono $N/2$ (per N pari) o $(N+1)/2$ (per N dispari) moltiplicazioni, invece delle N moltiplicazioni richieste nel caso generale mostrato in fig. 4.21. Tali reti sono riportate in fig. 4.24 per N pari ed in fig. 4.25 per N dispari. Forme trasposte possono naturalmente essere ricavate dalle fig. 4.24 e 4.25, nel modo già visto prima.

Imporre la condizione di simmetria (4.40) per i coefficienti del polinomio $H(z)$ comporta che gli zeri di $H(z)$ si verifichino in coppie speculari,

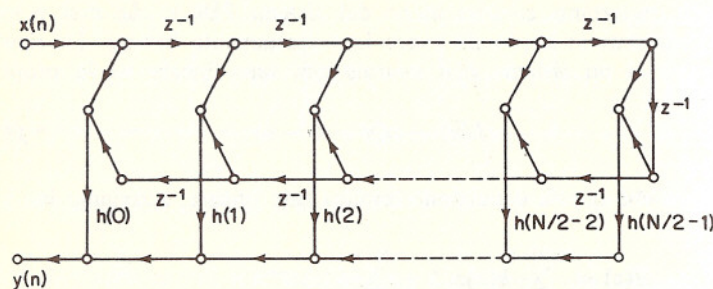


Fig. 4.24 Realizzazione in forma diretta di un sistema FIR di ordine pari con fase lineare.

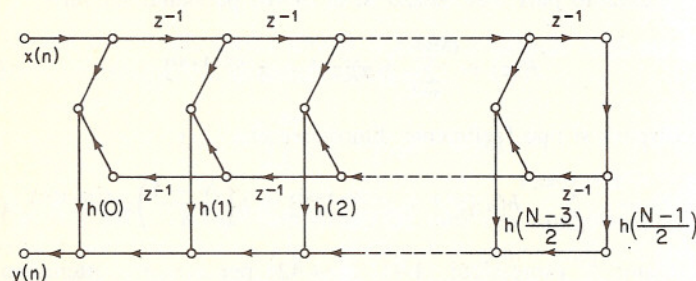


Fig. 4.25 Realizzazione in forma diretta di un sistema FIR di ordine dispari con fase lineare.

cioè, se z_0 è uno zero di $H(z)$, allora è zero di $H(z)$ anche $1/z_0$ (si veda il probl. 15 del cap. 2). Inoltre, se i coefficienti $h(n)$ sono reali, gli zeri di $H(z)$ saranno disposti in coppie complesse coniugate. Come conseguenza, gli zeri che non sono sul circolo unitario sono disposti in coppie reciproche. Gli zeri complessi che non sono sul circolo unitario sono disposti in gruppi di quattro, che corrispondono ai complessi coniugati ed ai reciproci. Se uno zero è sul circolo unitario, il suo reciproco è anche il suo coniugato; di conseguenza, è conveniente raggruppare in coppie gli zeri complessi sul circolo unitario. Gli zeri reali sul circolo unitario, cioè gli zeri per $z = 1$ o per $z = -1$, sono reciproci e complessi coniugati di se stessi e di conseguenza debbono essere considerati individualmente. I quattro casi sono riassunti in fig. 4.26, dove gli zeri in z_1 , z_1^* , $1/z_1$ e $1/z_1^*$ sono considerati come un gruppo di quattro; gli zeri in z_2 e $1/z_2$ sono considerati come un gruppo di due, e così z_3 e z_3^* ; lo zero in z_4 va considerato da solo. In corrispondenza a questo raggruppamento di zeri, $H(z)$ può essere scomposto nel prodotto di fattori del primo, secondo e quarto ordine. Ognuno di questi fattori è un polinomio i cui

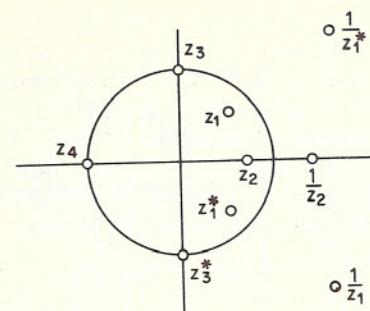


Fig. 4.26 Simmetria degli zeri di un sistema FIR a fase lineare.

coefficienti hanno la stessa simmetria di quelli di $H(z)$, cioè ogni fattore è un polinomio a fase lineare. Perciò possiamo ottenere una realizzazione espressa come la cascata di più sistemi a fase lineare del primo, secondo e quarto ordine. I sistemi del primo ordine corrispondono ad uno zero in $z = \pm 1$ e non richiedono moltiplicazioni. I fattori del secondo ordine saranno della forma

$$1 + az^{-1} + z^{-2}$$

e perciò avranno bisogno di un'unica moltiplicazione. Le sezioni del quarto ordine avranno la forma

$$a + bz^{-1} + cz^{-2} + bz^{-3} + az^{-4}$$

e richiederanno tre moltiplicazioni se realizzate come strutture a fase lineare come nel caso di fig. 4.25.

4.5.4 Struttura basata sul campionamento in frequenza

Nel cap. 3 abbiamo mostrato che la trasformata z di una sequenza di durata finita N può essere rappresentata mediante N campioni equispaziati sul circolo unitario. Per un filtro FIR, dalla relazione (3.18) consegue che la funzione di trasferimento può essere espressa come

$$H(z) = (1 - z^{-N}) \frac{1}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - W_N^{-k} z^{-1}} \quad (4.43)$$

essendo

$$W_N^{-k} = e^{j(2\pi k/N)}$$

e

$$\tilde{H}(k) = H(W_N^{-k}) = |H(k)| e^{j\theta(k)} \quad (4.44)$$

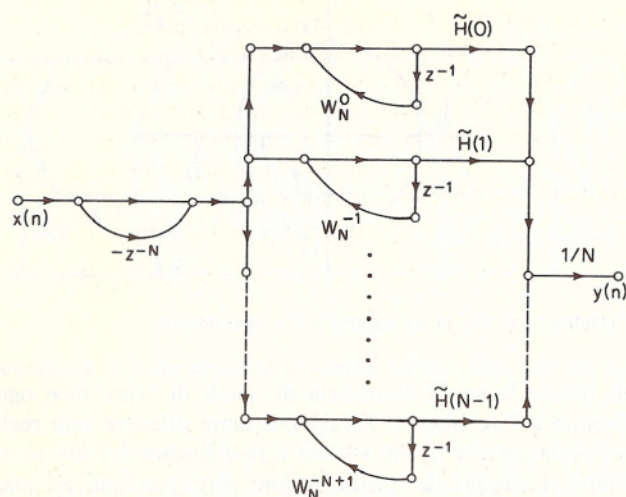


Fig. 4.27 Struttura basata sul campionamento in frequenza per realizzare un sistema FIR.

Le quantità $\tilde{H}(k)$ vengono chiamate *campioni in frequenza* poiché sono semplicemente campioni della risposta in frequenza del sistema¹.

L'espressione (4.43) suggerisce che un sistema FIR può essere realizzato come cascata di un sistema FIR molto semplice con un sistema IIR, come è mostrato nella fig. 4.27. La funzione di trasferimento del sistema FIR è $1 - z^{-N}$, e gli zeri di questo sistema si hanno per $z_k = \exp[j(2\pi/N)k]$. La parte IIR nella fig. 4.27 consiste in una combinazione in parallelo di N sistemi complessi del primo ordine con poli per $z_k = \exp[j(2\pi/N)k]$. Questi sistemi del primo ordine hanno poli esattamente sul circolo unitario, il cui effetto è proprio quello di cancellare gli zeri del sistema FIR. In pratica, i problemi di stabilità legati ai poli sul circolo unitario vengono superati campionando la funzione di trasferimento $H(z)$ su una circonferenza di raggio r , con r leggermente minore dell'unità [9]. In questo caso $H(z)$ può essere espresso come

$$H(z) = (1 - r^N z^{-N}) \frac{1}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - r W_N^{-k} z^{-1}} \quad (4.45)$$

dove la rappresentazione esatta di $H(z)$ richiede che sia

$$\tilde{H}(k) = H(r W_N^{-k}) \quad (4.46)$$

¹ Ricordiamo al lettore che, in accordo con la notazione del cap. 3, $\tilde{H}(k)$ è periodica con periodo N . $H(k)$, la DFT di $h(n)$, è un periodo della sequenza periodica $\tilde{H}(k)$.

In pratica, tuttavia, r viene scelto prossimo all'unità cosicché l'uso della (4.44) al posto della (4.46) produce un errore piccolo.

In generale, i campioni in frequenza $\tilde{H}(k)$ sono complessi e tali sono anche le quantità W_N^{-k} . Perciò, la realizzazione di un sistema FIR secondo lo schema di fig. 4.27 richiede operazioni di aritmetica complessa. Tuttavia, come sappiamo dal cap. 3, se i campioni della risposta all'impulso $h(n)$ sono reali, i campioni in frequenza espressi in forma polare soddisfano le seguenti condizioni di simmetria:

$$\begin{aligned} |\tilde{H}(k)| &= |\tilde{H}(N-k)| \\ \theta(k) &= -\theta(N-k), \quad k = 0, 1, \dots, N-1 \end{aligned} \quad (4.47)$$

dove, se $\tilde{H}(0) > 0$, si ha

$$\theta(0) = 0$$

Inoltre, poiché $(W_N^{-k})^* = W_N^{-(N-k)}$, le reti del primo ordine rappresentate nella fig. 4.27 compaiono in coppie complesse coniugate eccetto la rete con il polo in W_N^0 e, per N pari, la rete col polo in $W_N^{-N/2}$. Di conseguenza, le reti complesse del primo ordine possono essere raggruppate in coppie complesse coniugate e realizzate come reti del secondo ordine con coefficienti reali. In particolare, supponendo che N sia pari, possiamo esprimere la (4.45) come

$$H(z) = \frac{1 - r^N z^{-N}}{N} \left[\sum_{k=1}^{(N/2)-1} \frac{\tilde{H}(k)}{1 - r W_N^{-k} z^{-1}} + \sum_{k=N/2+1}^{N-1} \frac{\tilde{H}(k)}{1 - r W_N^{-k} z^{-1}} + \frac{\tilde{H}(0)}{1 - r z^{-1}} + \frac{\tilde{H}(N/2)}{1 + r z^{-1}} \right]$$

che diventa, cambiando l'indice di somma nella seconda sommatoria,

$$H(z) = \frac{1 - r^N z^{-N}}{N} \left[\sum_{k=1}^{(N/2)-1} \left(\frac{\tilde{H}(k)}{1 - r W_N^{-k} z^{-1}} + \frac{\tilde{H}(N-k)}{1 - r W_N^{-N+k} z^{-1}} \right) + \frac{\tilde{H}(0)}{1 - r z^{-1}} + \frac{\tilde{H}(N/2)}{1 + r z^{-1}} \right] \quad (4.48)$$

Usando la relazione (4.47) e il fatto che $(W_N^{-k})^* = W_N^{-(N-k)}$, possiamo riscrivere la (4.48) come segue

$$H(z) = (1 - r^N z^{-N}) \left[\sum_{k=1}^{(N/2)-1} \frac{2 |\tilde{H}(k)|}{N} H_k(z) + \frac{\tilde{H}(0)/N}{1 - r z^{-1}} + \frac{\tilde{H}(N/2)/N}{1 + r z^{-1}} \right] \quad (4.49)$$

dove

$$H_k(z) = \frac{\cos(\theta(k)) - r z^{-1} \cos[\theta(k) - 2\pi k/N]}{1 - 2r z^{-1} \cos(2\pi k/N) + r^2 z^{-2}} \quad (4.50)$$

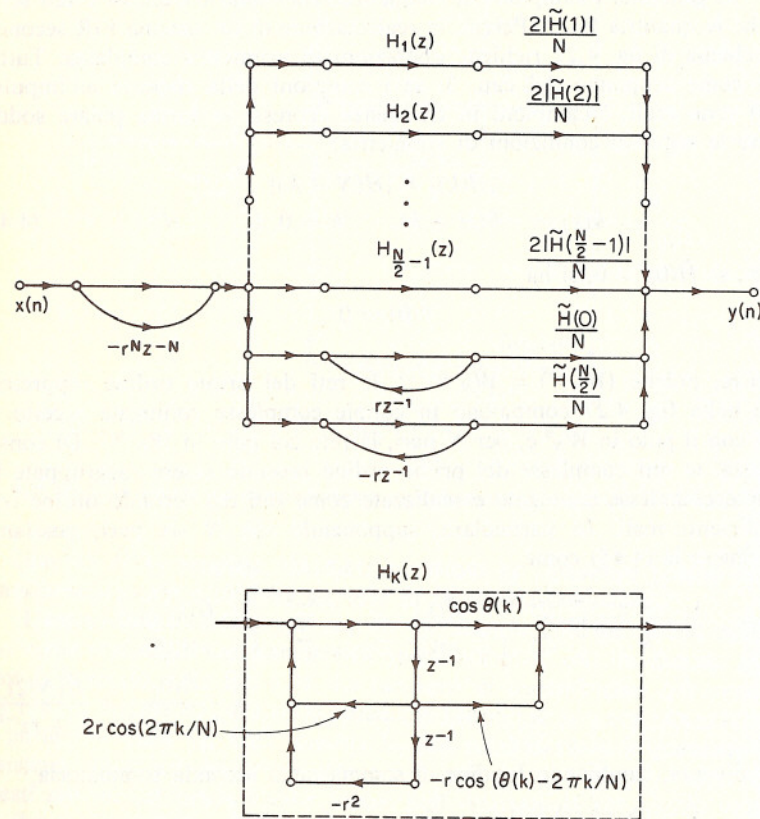


Fig. 4.28 Struttura basata sul campionamento in frequenza effettuato su un circolo prossimo a quello unitario e con i poli complessi realizzati in sezioni del secondo ordine.

Questa relazione suggerisce la struttura di rete rappresentata nella fig. 4.28, dove tutte le operazioni aritmetiche sono ora su numeri reali.

Quando N è dispari, non ci sarà un campione in frequenza per $k = N/2$. Perciò il termine contenente $|\tilde{H}(N/2)|$ mancherà dalla (4.49) e dalla fig. 4.28. Se il sistema ha fase lineare, le (4.49) e (4.50) possono essere ulteriormente semplificate (cfr. probl. 16). Una struttura simile può essere ricavata esprimendo $H(z)$ in termini di campioni distanziati di un angolo π/N dai precedenti campioni in frequenza. La derivazione di tale struttura è trattata, a grandi linee, nel probl. 17 di questo capitolo.

I principali vantaggi delle realizzazioni basate sul campionamento in frequenza sono due. In primo luogo, notiamo che i termini che moltiplicano le uscite di ciascun sistema del secondo ordine di fig. 4.28 sono proporzionali a campioni della risposta in frequenza equispaziati sul circolo unitario. Questi valori si possono ovviamente ottenere dalla DFT della risposta all'impulso. Se il filtro da realizzare è un filtro selettivo in frequenza con una o più bande oscure, esso può essere progettato, come verrà discusso nel cap. 5, in maniera tale che i campioni in frequenza nelle bande oscure siano zero, riducendo così il numero di sistemi del secondo ordine $H_k(z)$ che devono essere realizzati. Se la maggioranza dei campioni in frequenza sono zero, come nel caso di un filtro passa-basso o passa-banda a banda stretta, allora una struttura basata sul campionamento in frequenza può richiedere meno moltiplicazioni di una realizzazione in forma diretta. Naturalmente, la realizzazione basata sul campionamento in frequenza richiederà sempre più memoria di una realizzazione in forma diretta.

Un secondo vantaggio deriva dall'osservazione che i poli e gli zeri della struttura del filtro dipendono soltanto dalla lunghezza della risposta all'impulso. Se un ingresso deve essere filtrato con un banco di filtri FIR (cioè con un certo numero di differenti risposte all'impulso di lunghezza N), allora un'unica realizzazione del fattore $(1 - z^{-N})$, e di ogni sezione del secondo ordine, servirà per tutti i filtri. Inoltre, la struttura della fig. 4.28 è molto modulare, prestandosi quindi a moltiplicazione nel tempo delle sezioni del secondo ordine.

4.5.5 Strutture basate su formule di interpolazione polinomiale

Come abbiamo notato, la funzione di trasferimento di un sistema FIR è un polinomio di grado $N-1$ nella variabile z^{-1} , dove N è la lunghezza della risposta all'impulso. È ben noto che un polinomio di grado $N-1$ è univocamente determinato dai suoi valori in N punti distinti. Ci sono diverse formule di interpolazione polinomiale, come le formule di Lagrange e di Newton, che specificano un polinomio in termini di N valori. In modo alternativo, possiamo specificare il valore di un polinomio e delle sue prime N derivate per un particolare valore di z^{-1} e costruire il polinomio dalla sua rappresentazione in serie di Taylor. Schuessler [10] ha dimostrato che tali rappresentazioni della funzione di trasferimento danno luogo a particolari strutture per la realizzazione di sistemi FIR.

La struttura basata sul campionamento in frequenza del paragrafo precedente costituisce un esempio in cui il polinomio che rappresenta $H(z)$ è costruito per mezzo di interpolazione trigonometrica tra punti equispaziati sul circolo unitario. Dalla relazione (4.43) si può facilmente

dedurre una generalizzazione della struttura basata sul campionamento in frequenza, chiamata struttura di Lagrange. In primo luogo, notiamo che la (4.43) può essere espressa come

$$H(z) = P(z) \sum_{k=0}^{N-1} \frac{H(z_k)}{N(1 - z_k z^{-1})} \quad (4.51)$$

in cui è

$$P(z) = \prod_{k=0}^{N-1} (1 - z_k z^{-1}) = 1 - z^{-N} \quad (4.52)$$

e

$$z_k = e^{j(2\pi/N)k}$$

Si può far vedere facilmente che le (4.51) e (4.52) corrispondono ad un polinomio di grado $N-1$ in z^{-1} e che questo polinomio dà valori corretti nei punti di campionamento (cfr. il probl. 14 di questo capitolo).

La (4.51) è una forma particolare della formula di interpolazione di Lagrange relativa al caso in cui i punti di campionamento sono equispaziati sul circolo unitario. In generale, i punti di campionamento z_k possono essere scelti arbitrariamente nel piano z . In tal caso si ha la seguente rappresentazione di $H(z)$:

$$H(z) = P(z) \sum_{k=0}^{N-1} \frac{H(z_k)}{P_k(z_k)(1 - z_k z^{-1})} \quad (4.53)$$

in cui è

$$P(z) = \prod_{k=0}^{N-1} (1 - z_k z^{-1}) \quad (4.54)$$

e

$$P_k(z) = \prod_{\substack{i=0 \\ i \neq k}}^{N-1} (1 - z_i z^{-1}) \quad (4.55)$$

Si può verificare facilmente che la (4.53) rappresenta un polinomio di grado $N-1$ e fornisce il valore corretto nei punti di campionamento z_k . Essa costituisce perciò una rappresentazione appropriata della funzione di trasferimento di un sistema FIR. La struttura di rete suggerita dalla (4.53) è rappresentata nella fig. 4.29. In particolare otteniamo una struttura molto simile alla fig. 4.27. Se scegliamo punti di campionamento che sono coppie complesse coniugate, e se la risposta all'impulso $h(n)$ è reale, allora nella sommatoria della (4.53) si possono combinare i termini com-

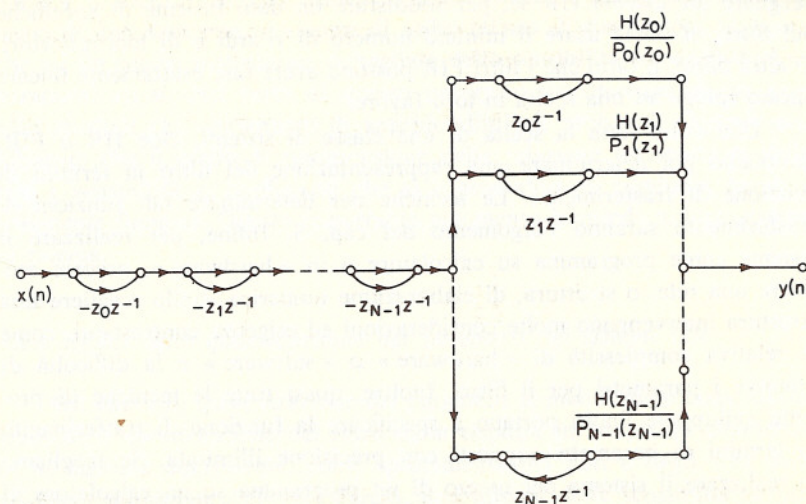


Fig. 4.29 Struttura di Lagrange per la realizzazione di un sistema FIR.

pletti coniugati in fattori del secondo ordine proprio come nel caso del campionamento in frequenza, ottenendo una rete simile alla fig. 4.28. Inoltre, possiamo associare tra loro fattori del polinomio $P(z)$ o in coppie complesse o tutti insieme per ottenere o una realizzazione in forma a cascata o una realizzazione in forma diretta della parte non ricorsiva della struttura.

Schuessler [10] ha discusso altre strutture di reti FIR basate sulle formule di interpolazione di Newton e di Hermite e sulle espansioni in serie di Taylor di $H(z)$. Tutte queste strutture, compresa la struttura basata sul campionamento in frequenza, in generale richiedono più moltiplicazioni e ritardi sia della forma diretta che della forma in cascata. Perciò, l'utilità di tali strutture consiste solo in possibili vantaggi rispetto alla sensibilità ad effetti di quantizzazione e nell'adattare metodi di progetto e tecniche di realizzazione del sistema.

4.6 EFFETTI DELLA QUANTIZZAZIONE DEI PARAMETRI

I sistemi lineari invarianti alla traslazione vengono spesso usati per attuare modifiche dello spettro di un segnale o altre operazioni di filtraggio. Nel progetto di tali sistemi sono importanti alcune considerazioni. In primo luogo occorre scegliere tra un sistema IIR e un sistema FIR. Questa scelta può dipendere da molti fattori. Per esempio, può convenire

scegliere un sistema IIR se, per soddisfare un dato insieme di specifiche sul filtro, si vuole usare il minimo numero di ritardi e di moltiplicatori. D'altra parte, il fatto che i filtri FIR possono avere fase esattamente lineare spesso spinge ad una scelta in loro favore.

Una volta fatta la scelta di una classe di sistemi, cioè IIR o FIR, dobbiamo poi determinare una rappresentazione del filtro in termini di funzione di trasferimento. Le tecniche per determinare tali funzioni di trasferimento saranno l'argomento del cap. 5. Infine, per realizzare il sistema come programma su calcolatore o in « hardware », occorre scegliere una rete, o struttura, di elaborazione numerica. Nello scegliere una struttura intervengono molte considerazioni ed esigenze contrastanti, come la relativa complessità di « hardware » o « software » e la difficoltà di ottenere i parametri per il filtro. Inoltre, quasi tutte le tecniche di progetto sviluppate finora portano a specificare la funzione di trasferimento in termini di parametri supposti con precisione illimitata. Se scegliamo di realizzare il sistema per mezzo di un programma su un calcolatore di impiego generale, la precisione con cui i parametri possono essere specificati è in generale limitata dalla lunghezza di parola della memoria del calcolatore. In realizzazioni « hardware » è ovviamente desiderabile minimizzare la lunghezza dei registri che devono essere disponibili per memorizzare i parametri del filtro.

Poiché i parametri usati nel realizzare un dato filtro non saranno in generale esatti, i poli e gli zeri della funzione di trasferimento saranno generalmente differenti dai poli e zeri desiderati. Questo allontanarsi di poli e zeri (solo zeri nel caso FIR) dalle posizioni volute fa sì che la risposta in frequenza risultante differisca dalla risposta in frequenza desiderata e, se gli errori di quantizzazione dei coefficienti sono elevati, il sistema può non soddisfare le specifiche di progetto. Inoltre, nel caso IIR, uno o più poli possono spostarsi fuori del cerchio unitario, facendo sì che il sistema risultante diventi instabile, quindi inutile per l'applicazione desiderata. In generale, l'effetto della quantizzazione dei coefficienti dipende molto dalla struttura usata per realizzare il sistema.

Come abbiamo visto, quando i parametri di rete sono rappresentati con precisione infinita, c'è tutta una varietà di strutture di rete che realizzano una data funzione di trasferimento. C'è da aspettarsi che alcune di tali strutture siano meno sensibili di altre alla quantizzazione dei parametri: ciò vuol dire che la funzione di trasferimento della realizzazione sarà più vicina, in qualche senso, alla funzione di trasferimento desiderata. Sfortunatamente, non è ancora stato sviluppato nessun metodo sistematico per determinare la migliore realizzazione date le condizioni su numero di moltiplicatori, lunghezza di parola, e numero di ritardi. In pratica, la scelta è generalmente limitata alle forme di rete dei par. 4.3,

4.4 e 4.5. Inoltre, per determinare la quantizzazione accettabile dei parametri di una data rete, si ricorre di solito all'uso di simulazioni piuttosto che ad un'analisi matematica dettagliata della sensibilità (o sensitività) ai parametri stessi. Nel resto di questo paragrafo discuteremo alcuni semplici risultati che consentono di approfondire tali problemi. Occorre sottolineare che il grado di comprensione della relazione tra struttura della rete e sensibilità ai coefficienti che si è generalmente raggiunto finora è veramente limitato. Questo importante argomento è quindi tuttora oggetto di ricerca.

4.6.1 Effetti della quantizzazione dei parametri nei sistemi IIR

Gli effetti di quantizzazione dei parametri si manifestano come deviazioni delle caratteristiche del filtro dalla risposta in frequenza desiderata o, in maniera equivalente, come spostamento dei poli e degli zeri dalle posizioni desiderate [11-13]. Perciò, una misura della sensibilità di una data realizzazione di una rete alla quantizzazione dei parametri è l'errore nella posizione di poli e zeri causato da un errore nei moltiplicatori della rete.

Per indicare come la quantizzazione dei parametri influenza le posizioni dei poli e degli zeri, consideriamo la funzione di trasferimento espressa come

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (4.56)$$

I coefficienti a_k e b_k sono i coefficienti desiderati in una realizzazione in forma diretta del sistema. Con coefficienti quantizzati realizziamo di fatto un sistema la cui funzione di trasferimento è

$$\hat{H}(z) = \frac{\sum_{k=0}^M \hat{b}_k z^{-k}}{1 - \sum_{k=1}^N \hat{a}_k z^{-k}} \quad (4.57)$$

dove

$$\begin{aligned} \hat{a}_k &= a_k + \Delta a_k \\ \hat{b}_k &= b_k + \Delta b_k \end{aligned}$$

Si supponga che i poli di $H(z)$ siano situati nelle posizioni $z = z_i$, $i = 1, 2, \dots, N$; cioè che il polinomio denominatore della funzione di trasferimento, espresso sotto forma di fattori, sia

$$P(z) = 1 - \sum_{k=1}^N a_k z^{-k} = \prod_{k=1}^N (1 - z_k z^{-1}) \quad (4.58)$$

Inoltre, definiamo i poli di $\hat{H}(z)$ come $z_i + \Delta z_i$, con $i = 1, 2, \dots, N$. Si può esprimere l'errore Δz_i in funzione degli errori nei coefficienti come

$$\Delta z_i = \sum_{k=1}^N \frac{\partial z_i}{\partial a_k} \Delta a_k, \quad i = 1, 2, \dots, N \quad (4.59)$$

Usando la relazione (4.5) e il fatto che è

$$\left(\frac{\partial P(z)}{\partial z_i} \right)_{z=z_i} \frac{\partial z_i}{\partial a_k} = \left(\frac{\partial P(z)}{\partial a_k} \right)_{z=z_i}$$

segue che

$$\frac{\partial z_i}{\partial a_k} = \frac{z_i^{N-k-1}}{\prod_{\substack{l=1 \\ l \neq i}}^N (z_i - z_l)} \quad (4.60)$$

La relazione (4.60) è una misura della sensibilità del polo i -mo ad un cambiamento (errore) nel coefficiente k -mo del polinomio denominatore di $H(z)$ [questo risultato è valido solo per poli semplici, come è evidente dalla relazione (4.58). L'estensione a poli multipli è immediata]. Poiché per la forma diretta gli zeri dipendono solo dai coefficienti b_k del numeratore, si può ottenere un risultato completamente analogo per la sensibilità degli zeri ad errori nei b_k .

Un risultato di questa forma fu usato da Kaiser [12, 13] per primo per dimostrare che, se i poli (o zeri) sono fittamente raggruppati, è possibile che piccoli errori nei coefficienti causino grandi spostamenti dei poli (o zeri). Ciò si può vedere considerando il denominatore della (4.60). Ogni fattore $(z_i - z_l)$ può essere rappresentato come un vettore nel piano z come mostrato nella fig. 4.30. Il modulo del denominatore della (4.60) è uguale al prodotto delle lunghezze dei vettori da tutti i poli rimanenti al polo z_i . Perciò, se i poli (o zeri) sono fittamente raggruppati come in fig. 4.30 (a), che corrisponde a un filtro passa-banda a banda stretta, o come in fig. 4.30 (b), che corrisponde a un filtro passa-basso a banda stretta, allora ci si può aspettare che i poli della realizzazione in forma diretta siano piuttosto sensibili ad errori di quantizzazione nei coefficienti. Inoltre, è evidente che quanto maggiore è il numero delle radici, tanto maggiore è la sensibilità.

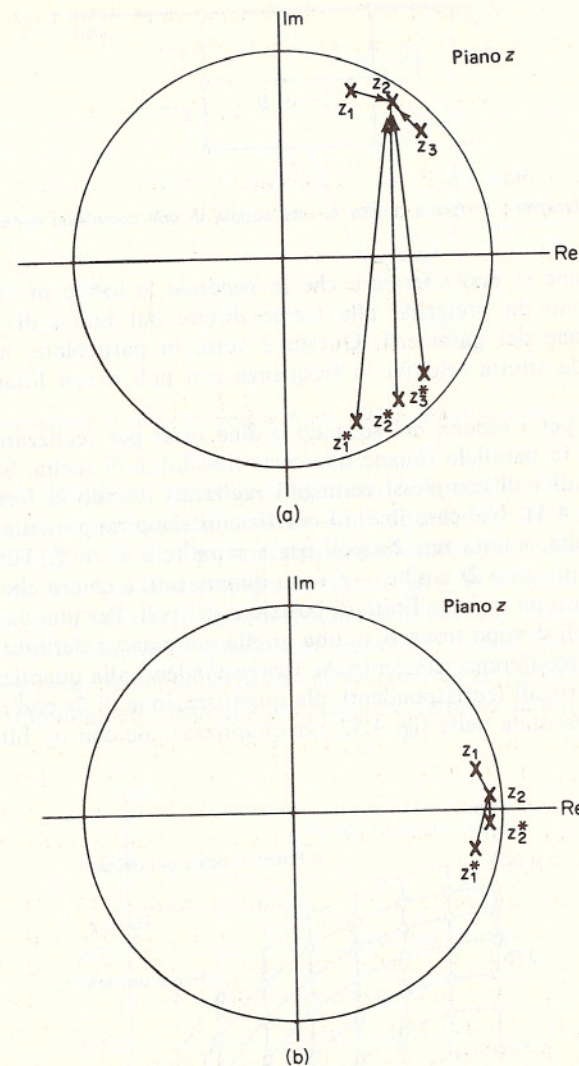


Fig. 4.30 Rappresentazione dei fattori della (4.58) come vettori nel piano z : (a) filtro passa-banda a banda stretta; (b) filtro passa-basso a banda stretta.

Le forme in cascata e in parallelo, d'altra parte, realizzano ciascuna coppia di poli complessi coniugati separatamente. Perciò l'errore in un dato polo è indipendente dalla sua distanza dagli altri poli del sistema. Per

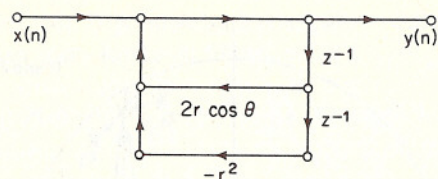


Fig. 4.31 Realizzazione in forma diretta di una coppia di poli complessi coniugati.

questa ragione si può affermare che in generale le forme in cascata e in parallelo sono da preferirsi alle forme dirette dal punto di vista della quantizzazione dei parametri. Questo è vero, in particolare, nel caso di filtri a banda stretta selettivi in frequenza con poli e zeri fittamente raggruppati.

Anche per i sistemi del secondo ordine usati per realizzare le forme in cascata e in parallelo rimane una certa flessibilità di scelta. Si consideri una coppia di poli complessi coniugati realizzati usando la forma diretta come in fig. 4.31. Nel caso in cui i coefficienti siano rappresentati con precisione infinita, questa rete ha poli per $z = re^{j\theta}$ e $z = re^{-j\theta}$. Tuttavia, se i fattori moltiplicativi $2r \cos \theta$ e $-r^2$ sono quantizzati, è chiaro che si può ottenere soltanto un insieme finito di posizioni dei poli. Per una data quantizzazione, i poli devono trovarsi su una griglia nel piano z definita dall'intersezione di circonferenze concentriche (corrispondenti alla quantizzazione di r^2) e rette verticali (corrispondenti alla quantizzazione di $2r \cos \theta$). Tale griglia è rappresentata nella fig. 4.32 per quantizzazione con tre bit, cioè r^2 e

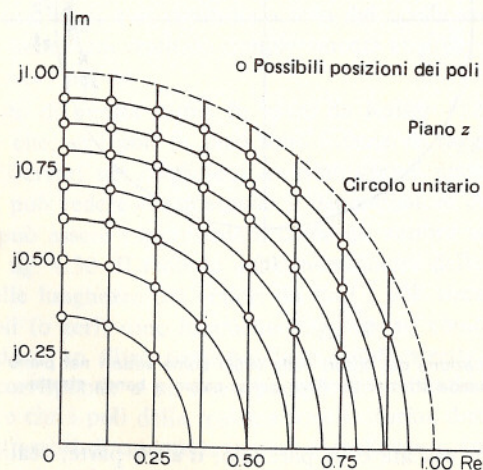


Fig. 4.32 Griglia di possibili posizioni dei poli per la rete della fig. 4.31 quando r^2 e $2r \cos \theta$ sono quantizzati con tre bit.

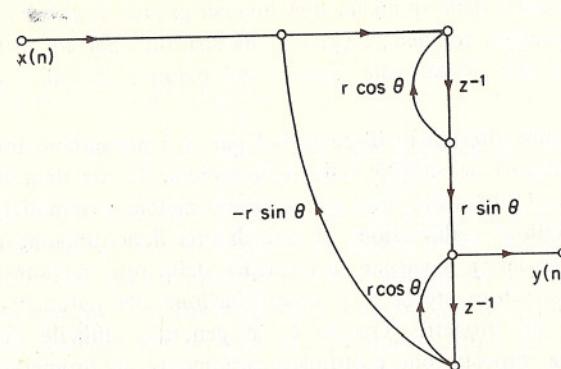


Fig. 4.33 Realizzazione in forma accoppiata di una coppia di poli complessi coniugati.

$2r \cos \theta$ sono limitati ad otto valori differenti. Una realizzazione alternativa è quella in forma accoppiata proposta da Gold e Rader [9], come rappresentato nella fig. 4.33. Si dimostra facilmente (v. probl. 1 di questo capitolo) che le funzioni di trasferimento per le reti di fig. 4.31 e fig. 4.33 hanno gli stessi poli per coefficienti con precisione infinita. Per realizzare la rete della fig. 4.33 dobbiamo quantizzare $r \cos \theta$ ed $r \sin \theta$. Perciò possiamo ottenere l'insieme finito delle posizioni dei poli come rappresentato nella fig. 4.34. Chiaramente, possiamo ottenere un numero molto maggiore di

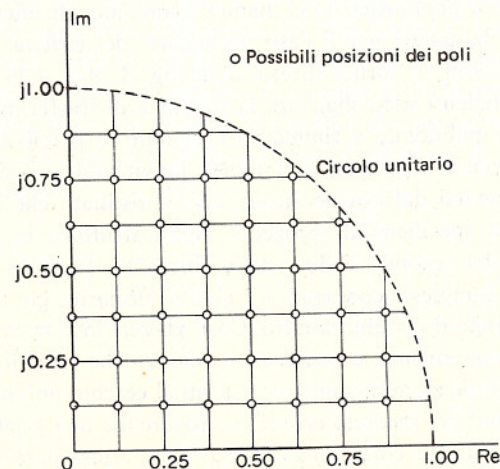


Fig. 4.34 Griglia di possibili posizioni dei poli per la rete di fig. 4.33 quando i coefficienti $r \cos \theta$ e $r \sin \theta$ sono quantizzati con tre bit.

strutture, ciascuna delle quali ha una diversa griglia di possibili posizioni dei poli. In pratica, conviene scegliere una struttura per cui la griglia corrispondente è più densa nella regione del piano z in cui si desiderano i poli.

Le strutture alternative discusse nel par. 4.3 presentano interesse per la possibile minore sensibilità nella realizzazione di una data funzione di trasferimento. Attualmente, non si conoscono metodi sistematici per determinare la migliore realizzazione di rete di una data funzione di trasferimento. In alternativa a variare la struttura della rete, si potrebbe tenere in conto la struttura stessa e la quantizzazione dei parametri già nella fase iniziale del progetto. Questo è, in generale, difficile da ottenere, benché questa impostazione costituisca certamente un promettente campo di ricerca.

4.6.2. Effetti della quantizzazione dei parametri nei sistemi FIR

Ricordiamo che per un sistema FIR dobbiamo preoccuparci solo della posizione degli zeri della funzione di trasferimento, in quanto tutti i poli, per un sistema FIR causale, cadono in $z = 0$. È chiaro che per la realizzazione in forma diretta del par. 4.5 si potrebbero derivare espressioni simili alla (4.60) relative all'errore in uno zero dovuto a errori nei coefficienti (risposta all'impulso). La conclusione è la stessa: cioè, in generale, si ha maggior controllo sulla posizione degli zeri in una realizzazione in cascata che non in una realizzazione in forma diretta.

Herrmann e Schuessler [14] hanno esaminato gli effetti della quantizzazione dei parametri per il caso particolare dei sistemi a fase lineare. Per le realizzazioni in forma diretta delle fig. 4.24 e 4.25 è chiaro che, anche se i coefficienti sono sbagliati, la funzione di trasferimento risultante sarà ancora un polinomio a simmetria speculare, e perciò il sistema avrà ancora fase lineare. Tuttavia, zeri molto ravvicinati sul circolo unitario possono allontanarsi dal circolo stesso, con il risultato che il sistema non soddisfa più le specifiche di progetto. Nella struttura in cascata, se si usano sezioni del secondo ordine della forma $(1 + az^{-1} + z^{-2})$ per ogni coppia di zeri complessi coniugati sul circolo unitario, gli zeri si possono spostare solo lungo il circolo unitario. Così, gli zeri in $z = \pm 1$ possono essere realizzati esattamente, e zeri reali sia dentro che fuori il cerchio unitario restano reali. Se gli zeri complessi esterni al cerchio unitario fossero realizzati con sezioni del secondo ordine, si troverebbe una griglia di possibili posizioni per ogni zero come quella di fig. 4.32, estesa però al di fuori del cerchio unitario. Se si vuole mantenere la fase lineare, occorre assicurarsi che per ogni zero interno al cerchio unitario esista uno zero coniugato reci-

proco esterno al cerchio. Si può ottenere questo esprimendo il fattore del quarto ordine corrispondente agli zeri in $z = re^{\pm j\theta}$ e $z = (1/r)e^{\pm j\theta}$ come

$$1 + d_1 z^{-1} + d_0 z^{-2} + d_1 z^{-3} + z^{-4} = \frac{1}{r^2} (1 - 2r \cos \theta z^{-1} + r^2 z^{-2})(r^2 - 2r \cos \theta z^{-1} + z^{-2}) \quad (4.61)$$

La sottorete corrispondente è mostrata in fig. 4.35. La griglia delle posizioni possibili per gli zeri è riportata di nuovo in fig. 4.36, questa volta per quantizzazione con cinque bit (32 valori diversi per r^2 e $2r \cos \theta$).

La realizzazione introdotta richiede chiaramente più moltiplicatori di quanti siano necessari per un sistema del quarto ordine. Se le sezioni del quarto ordine sono realizzate nella forma diretta con fase lineare, si ottiene la rete di fig. 4.37. La griglia delle posizioni possibili degli zeri per la forma diretta è mostrata in fig. 4.38. Mentre l'uso di sezioni del quarto ordine nella realizzazione di filtri FIR a fase lineare offre il vantaggio di mantenere le caratteristiche di linearità di fase del filtro indipendentemente dalla quantizzazione dei coefficienti, è stato mostrato [14-16] che per molti filtri le caratteristiche risultanti presentano un'estrema sensibilità alla quantizzazione. Ne segue che spesso è meglio realizzare filtri FIR a fase lineare come cascata di sezioni del secondo ordine o in forma diretta.

La discussione precedente sulla sensibilità rispetto alla quantizzazione dei parametri è stata basata soprattutto sull'esame della sensibilità delle posizioni dei poli e degli zeri. È anche possibile analizzare le strutture in base ad altre considerazioni di sensibilità. Anche se questo settore rimane oggetto di attività di ricerca, alcuni utili teoremi sulla sensibilità di rete possono ricavarsi sulla base della teoria dei grafi lineari di flusso di segnale. Chiudiamo questo capitolo con la presentazione di un interessante e utile teorema sui grafi di flusso di segnale, il teorema di Tellegen. Conseguenze di questo teorema sono, tra le altre, il teorema di trasposizione per i filtri numerici, citato nel par. 4.4, e un insieme di relazioni di sensibilità che vedremo nel seguito.

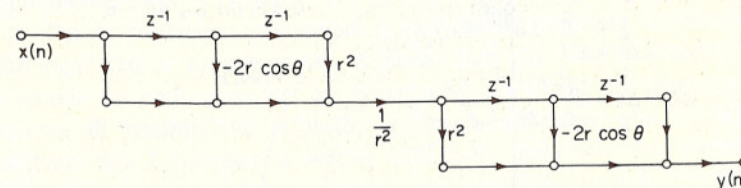


Fig. 4.35 Sottorete per la realizzazione dei fattori del quarto ordine in un sistema FIR a fase lineare in modo che sia mantenuta la fase lineare indipendentemente dalla quantizzazione dei parametri.

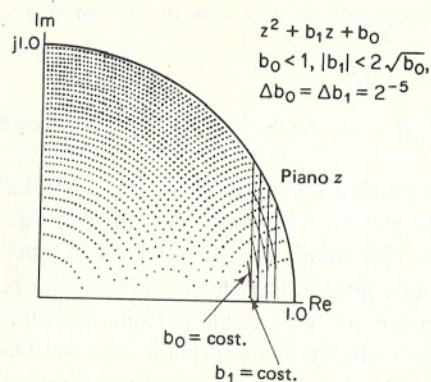


Fig. 4.36 Griglia delle possibili posizioni per gli zeri della sottorete di fig. 4.35 quando i coefficienti r^2 e $2r\cos\theta$ sono quantizzati con cinque bit (da Herrmann e Schuessler [14]).

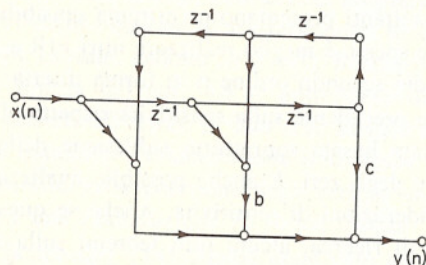


Fig. 4.37 Realizzazione in forma diretta con fase lineare per fattori del quarto ordine in un sistema FIR.

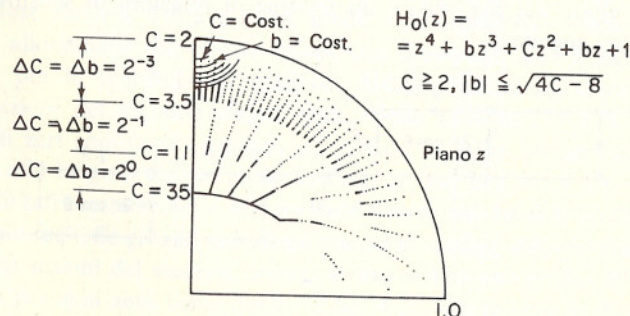


Fig. 4.38 Griglia delle possibili posizioni per gli zeri della sottorete di fig. 4.37 quando i coefficienti b e c sono quantizzati con cinque bit (da Herrmann e Schuessler [14]).

4.7 IL TEOREMA DI TELLEGEN PER I FILTRI NUMERICI E SUE APPLICAZIONI

Il teorema di Tellegen è un importante teorema fondamentale nella teoria classica delle reti. Il teorema è semplice e generale e parte della sua eleganza deriva dal fatto che molti teoremi di teoria delle reti, come la conservazione dell'energia, la reciprocità ecc., possono essere introdotti in maniera semplice come conseguenze del teorema di Tellegen.

Poiché le reti di elaborazione numerica non sono soggette alle leggi di Kirchhoff, non si può introdurre il teorema di Tellegen nella sua forma più generale. Se ne può però derivare una forma ristretta, la cosiddetta forma differenziale, in base alla quale dedurre poi molte utili proprietà delle reti stesse.

Una forma del teorema di Tellegen per i filtri numerici fu derivata per la prima volta da Seviara e Sablatash [18]. Una forma alternativa fu proposta da Fettweis [19]. La trattazione che faremo e la derivazione dei risultati seguono strettamente l'approccio di Fettweis. Nella presentazione del teorema di Tellegen è opportuno usare la notazione dei grafi di flusso di segnale sviluppata nel par. 4.1.

Il teorema di Tellegen per le reti classiche è espresso in forma di relazione tra la distribuzione delle tensioni in una rete e la distribuzione delle correnti in una seconda rete, dove l'unico legame tra le due reti è che esse hanno la stessa topologia. Analogamente consideriamo, per i grafi di flusso di segnale, due grafi di flusso aventi la stessa topologia e nessun'altra relazione. Avere la stessa topologia significa che esiste una corrispondenza biunivoca tra i nodi e i rami delle due reti (per quel che riguarda solo la posizione e la direzione). È però importante notare che il requisito di equivalenza topologica delle due reti è relativamente inessenziale. In particolare, se consideriamo ogni grafo di flusso come avente un ramo in entrambe le direzioni tra ogni coppia di nodi, con coefficiente di trasmissione nullo per alcuni rami, allora, presi due qualsiasi grafi di flusso con lo stesso numero di nodi, essi possono essere considerati topologicamente equivalenti.

Per la trattazione che segue è conveniente adottare la convenzione che ogni rete è rappresentata in modo tale che ogni nodo di rete ha associato un nodo sorgente, cui è connesso attraverso un ramo con coefficiente di trasmissione unitario. Inoltre, assumeremo che questo nodo sorgente non è collegato a nessun altro nodo della rete. La numerazione dei nodi sorgente sarà fatta coincidere con quella dei nodi di rete associati, in modo che la sorgente x_k sia l'ingresso al nodo di rete k . Se il nodo sorgente e il ramo corrispondente non sono rappresentati, questo implica che il valore del nodo sorgente è zero.

TEOREMA DI TELLEGEN. Consideriamo due grafi di flusso di segnale con la stessa topologia. Sia N il numero dei nodi della rete. Le variabili di nodo della rete, le uscite dei rami e i valori dei nodi sorgente sono indicati, rispettivamente, con w_k , v_{jk} e x_i nella prima rete e con w'_k , v'_{jk} e x'_i nella seconda. Allora risulta

$$\sum_{k=1}^N \sum_{j=1}^N (w'_k v_{jk} - w_k v'_{jk}) + \sum_{k=1}^N (w_k x'_k - w'_k x_k) = 0. \quad (4.62)$$

Dimostrazione. La dimostrazione della relazione (4.62) discende direttamente dalla definizione di grafo di flusso di segnale. Ricordiamo che le uscite dei rami sono legate alle variabili di nodo e agli ingressi nelle sorgenti dalla

$$w_k = \sum_{j=1}^N v_{jk} + \sum_{j=1}^M s_{jk}$$

e perciò, con la nostra convenzione sui nodi sorgente, si può scrivere

$$w_k = \sum_{j=1}^N v_{jk} + x_k \quad (4.63)$$

Partendo ora dall'identità

$$\sum_{k=1}^N (w_k w'_k - w'_k w_k) = 0 \quad (4.64)$$

la (4.62) si ricava immediatamente sostituendo la (4.63) nella (4.64).

La relazione (4.62) rappresenterà d'ora innanzi il teorema di Tellegen per grafi di flusso di segnale o, cosa equivalente, per le reti di elaborazione numerica. È importante notare che la sua derivazione dipende solo dalla relazione (4.63): non è quindi richiesto che il grafo di flusso sia lineare. Inoltre, è immediato mostrare che se le variabili w_k , w'_k , v_{jk} , v'_{jk} , x_k e x'_k sono ottenute per mezzo di un'operazione lineare rispettivamente da w_k , w'_k , v_{jk} , v'_{jk} , x_k e x'_k , allora risulta

$$\sum_{k=1}^N \sum_{j=1}^N (W_k V'_{jk} - W'_k V_{jk}) + \sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0 \quad (4.65)$$

Perciò il teorema di Tellegen vale sia per le sequenze [v. (4.62)] che per le trasformate z [v. (4.65)].

Nel resto della nostra trattazione ci occuperemo solo di grafi lineari di flusso di segnale che corrispondono a reti di elaborazione numerica. Per chiarire il concetto di topologie identiche e di variabili (e reti) contrassegnate o meno dall'apice, si considerino le due reti mostrate in fig. 4.39. Per sottolineare l'equivalenza della topologia abbiamo rappresentato con una linea tratteggiata un ramo con coefficiente di trasmissione nullo. Di solito, ovviamente, tali rami non vengono disegnati. La verifica della validità del teorema di Tellegen per queste due reti costituisce un semplice esercizio (v. probl. 18 di questo capitolo).

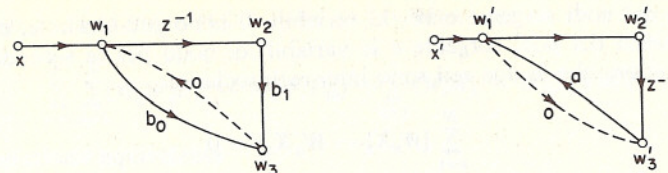


Fig. 4.39 Esempio di due reti che sono topologicamente equivalenti e perciò soddisfano il teorema di Tellegen.

4.7.1 Reti numeriche reciproche e inter-reciproche

Il concetto di reciprocità ha un ruolo importante nelle reti analogiche passive, costituite da resistenze, induttanze e capacità collegate tra loro. Per le reti numeriche esistono i concetti corrispondenti di reciprocità e inter-reciprocità. Per quanto riguarda la reciprocità, consideriamo una certa rete, eccitata da due diversi insiemi di sorgenti. Le trasformate z dei valori dei nodi sorgente per i due diversi insiemi saranno indicate con X_k e X'_k . Il valore della variabile di nodo per il nodo k -esimo quando la rete è eccitata dalle sorgenti senza apice sarà indicato con W_k . Quando la rete è eccitata dalle sorgenti contrassegnate dall'apice, questa variabile sarà indicata con W'_k . Allora si dice che la rete soddisfa la reciprocità se per qualsiasi coppia di insiemi di segnali di eccitazione risulta

$$\sum_{k=1}^N [W_k X'_k - W'_k X_k] = 0 \quad (4.66)$$

Per vedere una conseguenza della reciprocità, consideriamo due nodi qualsiasi a e b nella rete. Inoltre, supponiamo che tutti i valori dei nodi delle sorgenti senza apice siano zero eccetto X_a e che tutti quelli delle sorgenti con l'apice siano nulli all'infuori di X'_b . In più, assumiamo che sia $X_a = X'_b$. Allora dalla (4.66) segue che

$$W_b X'_b = W'_a X_a$$

ovvero

$$W_b = W'_a \quad (4.67)$$

In altri termini, come conseguenza della reciprocità, se eccitiamo il grafo nel nodo di rete a e osserviamo l'uscita al nodo b , allora, per un grafo reciproco, un'eccitazione uguale applicata al nodo b produce la stessa uscita al nodo a .

Per lo più le reti numeriche non sono reciproche. Un concetto collegato, che è più utile in questo caso, è quello di *inter-reciprocità*. Per definirlo consideriamo due distinti grafi di flusso di segnale. Siano X_k

i valori dei nodi sorgente e W_k le variabili di nodo per una rete, e X'_k e W'_k i valori dei nodi sorgente e le variabili di nodo per la seconda rete. Si dice allora che le due reti sono inter-reciproche se

$$\sum_{k=1}^N [W_k X'_k - W'_k X_k] = 0 \quad (4.68)$$

Questa equazione è formalmente identica alla (4.66), ma è importante aver presente che per la reciprocità le due reti (con apice e senz'apice) differiscono solo nelle sorgenti, mentre per l'inter-reciprocità sia le sorgenti che i coefficienti di trasmissione dei rami possono essere diversi nelle due reti. Una rete reciproca è anche inter-reciproca rispetto a se stessa. D'altro canto, due reti possono essere inter-reciproche senza che nessuna delle due sia reciproca.

4.7.2 Dimostrazione del teorema di trasposizione

Un'utile proprietà delle reti numeriche è che esse sono inter-reciproche rispetto alle loro trasposte. Per dimostrarlo, ricordiamo dal par. 4.4 che la trasposizione di un grafo di flusso si ottiene invertendo la direzione di tutti i rami ma lasciando inalterati i coefficienti di trasmissione.

Consideriamo una rete dove W_k indica la variabile di nodo relativa al nodo k -esimo. Il coefficiente di trasmissione dal nodo j al nodo k è indicato con F_{jk} , cioè risulta

$$V_{jk} = F_{jk} W_j \quad (4.69)$$

Nella rete trasposta, in cui la variabile del k -esimo nodo è W'_k , e il coefficiente di trasmissione del ramo tra i nodi j e k è F'_{jk} , risulta

$$V'_{jk} = F'_{jk} W'_j \quad (4.70)$$

Per definizione di rete trasposta si ha

$$F'_{jk} = F_{kj}$$

Per dimostrare che una rete e la sua trasposta sono inter-reciproche, cioè per provare la validità della relazione (4.68) sotto queste condizioni, utilizziamo il fatto che una rete e la sua trasposta hanno la stessa topologia, per cui vale il teorema di Tellegen [v. (4.65)]. Perciò si ha

$$\sum_{j=1}^N \sum_{k=1}^N (W_k V'_{jk} - W'_k V_{jk}) + \sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0 \quad (4.71)$$

Sostituendo le (4.69) e (4.70) nella (4.71) otteniamo

$$\sum_{j=1}^N \sum_{k=1}^N (W_k W'_j F'_{jk} - W'_k W_j F_{jk}) + \sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0$$

o la relazione equivalente

$$\sum_{j=1}^N \sum_{k=1}^N W_k W'_j F'_{jk} - \sum_{j=1}^N \sum_{k=1}^N W'_k W_j F_{jk} + \sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0 \quad (4.72)$$

Scambiando gli indici di somma nella prima sommatoria doppia si ricava

$$\sum_{j=1}^N \sum_{k=1}^N (W'_k W_j F'_{kj} - W_k W'_j F_{jk}) + \sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0 \quad (4.73)$$

Adesso, essendo le reti con e senza apice l'una la trasposta dell'altra, è $F'_{jk} = F_{jk}$ e quindi la sommatoria doppia è nulla e si ha

$$\sum_{k=1}^N (W_k X'_k - W'_k X_k) = 0 \quad (4.74)$$

col che è dimostrato che una rete e la sua trasposta sono inter-reciproche.

Una conseguenza interessante di questo fatto è che, nel caso di reti con un solo ingresso e una sola uscita, una rete e la sua trasposta hanno la stessa funzione di trasferimento. Questo risultato può anche essere derivato dalla formula del guadagno di Mason [1] ed è stato utilizzato nei par. 4.3 e 4.4 come mezzo per ottenere nuove realizzazioni di una rete.

Per dimostrare questo risultato a partire dalla (4.74), consideriamo due nodi qualsiasi a e b . Inoltre, supponiamo che tutti i nodi sorgente della rete originale (senz'apice) siano nulli eccetto X_a e che tutti i nodi sorgente della rete trasposta (con apice) siano nulli all'infuori di X'_b . Allora discende dalla (4.74) che

$$W_b X'_b = W'_a X_a \quad (4.75)$$

Dall'uguaglianza (4.75) si vede che, se si applica un'uguale eccitazione al nodo a della rete originaria e al nodo b della rete trasposta, allora al nodo a della rete trasposta si osserva la stessa risposta del nodo b della rete originaria.

4.7.3 Formula per la sensitività di rete

Nel par. 4.6 abbiamo discusso il problema della sensibilità ai parametri riguardo agli spostamenti dei poli e degli zeri. In particolare abbiamo confrontato le reti in forma diretta con le realizzazioni in cascata e in parallelo. Per reti più complesse, non è altrettanto facile ottenere una rela-

zione generale tra la posizione di poli e zeri e i parametri della rete. Però possiamo, con l'aiuto del teorema di Tellegen, ottenere un'espressione generale per la sensitività della funzione di trasferimento di una data rete a cambiamenti dei parametri della rete. Formule di questo tipo sono utili, ad esempio, nell'analisi di reti numeriche assistita da calcolatore.

Nell'ambito di questa discussione è opportuno definire la funzione di trasferimento tra due nodi arbitrari a e b della rete. A questo scopo assumiamo che tutti i nodi sorgente abbiano valore zero eccetto quello collegato al nodo a . La variabile di nodo $W_b(z)$ è allora espressa come

$$W_b(z) = T_{ab}(z)X_a(z) \quad (4.76)$$

T_{ab} è quindi la funzione di trasferimento dal nodo a al nodo b . La sensitività di questa funzione di trasferimento a variazioni in un coefficiente di trasmissione di ramo $F_{nm}(z)$ è definita come

$$\frac{\partial T_{ab}(z)}{\partial F_{nm}(z)}$$

Fettweis [19] e Seviara e Sablatash [18] hanno dimostrato che

$$\frac{\partial T_{ab}}{\partial F_{nm}} = T_{an}T_{mb} \quad (4.77)$$

dove non abbiamo indicato esplicitamente la dipendenza funzionale da (z) , e dove T_{an} è la funzione di trasferimento dal nodo a al nodo n e T_{mb} quella dal nodo m al nodo b .

Per dimostrare questo risultato consideriamo le tre reti rappresentate in fig. 4.40. La rete originale, mostrata in fig. 4.40 (a), è contraddistinta dalle variabili senza apice ed è chiamata per comodità *rete senza apici*. La fig. 4.40 (b) mostra la rete trasposta di quella originaria, detta *rete con apici*, e la fig. 4.40 (c) rappresenta il sistema originale con una perturbazione ΔF_{nm} nel coefficiente di trasmissione di ramo F_{nm} . Questa rete viene chiamata *rete con due apici*. Assumiamo che ogni rete sia eccitata dalla stessa sorgente X , come mostrato in fig. 4.40.

Usando la relazione (4.65) per i sistemi con apice e con doppio apice, si ottiene

$$\sum_{k=1}^N \sum_{j=1}^N (W'_k V''_{jk} - W''_k V'_{jk}) + \sum_{k=1}^N (W'_k X''_k - W''_k X'_k) = 0 \quad (4.78)$$

Spezzando la sommatoria doppia in due e scambiando gli indici nella seconda sommatoria doppia risultante, si ha

$$\sum_{k=1}^N \sum_{j=1}^N (W'_k V''_{jk} - W''_j V'_{kj}) + \sum_{k=1}^N (W'_k X''_k - W''_k X'_k) = 0 \quad (4.79)$$

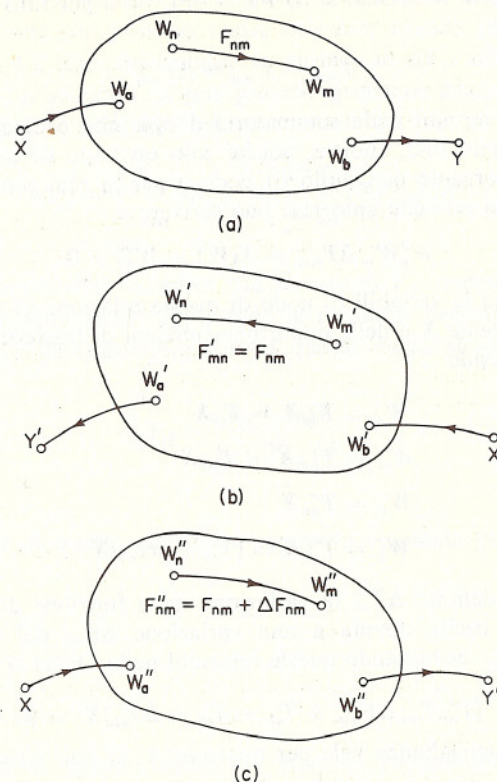


Fig. 4.40 Tre reti in relazione tra loro, usate per derivare la relazione di sensitività (4.77): (a) rete originaria; (b) rete trasposta (sistema con apici); (c) sistema originale con un errore nel coefficiente di trasmissione di ramo F_{nm} (sistema con due apici).

Usando la relazione (4.79) e il fatto che per definizione è

$$V''_{jk} = F''_{jk} W''_j$$

e

$$V'_{jk} = F'_{jk} W'_j$$

si ottiene

$$\sum_{k=1}^N \sum_{j=1}^N W''_j W'_k (F''_{jk} - F'_{kj}) + \sum_{k=1}^N (W'_k X''_k - W''_k X'_k) = 0 \quad (4.80)$$

Dalla fig. 4.40 osserviamo che le reti con apice e con doppio apice sono l'una la trasposta dell'altra ad eccezione del ramo (nm) . Perciò, in base

alla definizione di trasposta, si ha $F''_{jk} - F'_{kj} = 0$ per tutti i k e j escluso il ramo (nm) e

$$F''_{nm} - F'_{mn} = \Delta F_{nm}$$

Quindi tutti i termini della sommatoria doppia che compare nella (4.80) sono nulli eccetto uno. Inoltre, poiché solo un nodo di ciascuna rete ha un valore di sorgente non nullo (il nodo a per la rete con doppio apice, il nodo b per la rete con apice), si può scrivere

$$W''_a W'_m \Delta F_{nm} + X(W'_a - W''_b) = 0 \quad (4.81)$$

Ora esprimiamo le variabili di nodo di questa relazione in termini dell'ingresso alla sorgente X e delle opportune funzioni di trasferimento dai nodi sorgente, ottenendo

$$\begin{aligned} W'_a &= T'_{ba} X = T_{ab} X \\ W'_m &= T'_{bm} X = T_{mb} X \\ W''_n &= T''_{an} X \\ W''_b &= T''_{ab} X = [T_{ab} + \Delta T_{ab}] X \end{aligned}$$

dove abbiamo definito ΔT_{ab} la variazione nella funzione di trasferimento tra ingresso e uscita dovuta a una variazione ΔF_{nm} del coefficiente di trasmissione F_{nm} . Sostituendo queste relazioni nella (4.81) si ricava

$$[T''_{an} T_{mb} \Delta F_{nm} + T_{ab} - T_{ab} - \Delta T_{ab}] X^2 = 0$$

Poiché questa uguaglianza vale per qualsiasi X , si può scrivere

$$T''_{an} T_{mb} \Delta F_{nm} = \Delta T_{ab} \quad (4.82)$$

o anche

$$\frac{\Delta T_{ab}}{\Delta F_{nm}} = T''_{an} T_{mb} \quad (4.83)$$

Facendo il limite della (4.83) per $\Delta F_{nm} \rightarrow 0$ si ha

$$\lim_{\Delta F_{nm} \rightarrow 0} \left[\frac{\Delta T_{ab}}{\Delta F_{nm}} \right] = \lim_{\Delta F_{nm} \rightarrow 0} [T''_{an} T_{mb}] \quad (4.84)$$

Quando $\Delta F_{nm} \rightarrow 0$ la rete con doppio apice tende a quella senza apici, per cui risulta

$$\frac{\partial T_{ab}}{\partial F_{nm}} = T_{an} T_{mb} \quad (4.85)$$

Abbiamo quindi ricavato la relazione di sensitività desiderata, che lega la variazione nella funzione di trasferimento T_{ab} alla variazione nel coefficiente di trasmissione di ramo F_{nm} . La caratteristica interessante di questa

espressione è che la sensitività è espressa in termini di funzioni di trasferimento della rete che possono essere calcolate usando il metodo matriciale del par. 4.2. Per determinare la variazione ΔT_{ab} della funzione di trasferimento T_{ab} prodotta da una grande variazione ΔF_{nm} di F_{nm} , si può usare lo sviluppo in serie di Taylor

$$\Delta T_{ab} = \frac{\partial T_{ab}}{\partial F_{nm}} \Delta F_{nm} + \frac{1}{2} \frac{\partial^2 T_{ab}}{\partial F_{nm}^2} (\Delta F_{nm})^2 + \dots \quad (4.86)$$

Le derivate (sensitività) di ordine più elevato di T_{ab} rispetto a F_{nm} si possono calcolare dalla (4.85) applicando la regola di derivazione in catena. Crochiere [20] ha dimostrato che in questo modo si giunge all'espressione

$$\Delta T_{ab} = \frac{T_{an} T_{mb} \Delta F_{nm}}{1 - T_{mn} \Delta F_{nm}} \quad (4.87)$$

Nel probl. 19 di questo capitolo è suggerito un modo per derivare questo risultato.

SOMMARIO

L'oggetto di questo capitolo è stato la rappresentazione dei filtri numerici in termini di diagrammi a blocchi, grafi di flusso di segnale e notazione matriciale. Dopo avere introdotto da un punto di vista generale la rappresentazione mediante grafi di flusso di segnale e quella matriciale, abbiamo presentato un certo numero di strutture fondamentali per i filtri IIR e FIR.

Uno dei fattori importanti nella scelta di una struttura per la realizzazione di un filtro è l'effetto della precisione finita dei coefficienti. Per questa ragione abbiamo svolto alcune considerazioni sulla quantizzazione dei parametri in relazione alla scelta della struttura del filtro. Abbiamo poi introdotto il teorema di Tellegen per i filtri numerici. A partire da questo teorema sono state sviluppate parecchie proprietà importanti delle strutture dei filtri, compreso il teorema di trasposizione per i grafi di flusso di segnale. A conclusione del capitolo è stata derivata una formula di sensitività per le reti usando il teorema di Tellegen. Questa formula costituisce uno strumento utile per calcolare la sensitività delle strutture dei filtri e inoltre conduce a un'espressione per la sensitività a grandi variazioni dei parametri.

gli elementi della diagonale principale siano nulli. In questo problema si vuole dimostrare che l'affermazione è vera.

Siano c_{jk} gli elementi della matrice F .

- (a) Si assuma che nella matrice F ci siano soltanto zeri sopra la diagonale principale e che gli elementi della diagonale principale siano nulli, in modo che $c_{jk} = 0$ per $j \geq k$. Mostrare che in questo caso le variabili di nodo $w_1(n)$, $w_2(n)$, ... possono essere calcolate in sequenza, cioè, prima $w_1(n)$, poi $w_2(n)$, etc. Per far ciò è necessario mostrare che il calcolo di $w_i(n)$ non utilizza le variabili di nodo $w_l(n)$ con $l \geq i$. Il calcolo può, naturalmente, utilizzare qualsiasi elemento del vettore $w(n-1)$ in quanto questo corrisponde alla precedente iterazione del grafo di flusso.

La parte (a) mostra che la condizione trovata è una condizione sufficiente per la calcolabilità. Si vuole ora mostrare che essa è anche una condizione necessaria.

- (b) Si assuma che nella matrice F ci sia almeno un elemento non nullo sopra o sulla diagonale principale. Mostrare che in questo caso le variabili di nodo non possono essere calcolate in sequenza.

La parte (b) mostra che la condizione enunciata è una condizione necessaria poiché, se i nodi non possono essere numerati in modo tale che F sia nulla sopra e sulla diagonale principale, allora non c'è alcun ordinamento in corrispondenza del quale le variabili di nodo possano essere calcolate.

4. Nel derivare la rappresentazione matriciale delle eq. (4.29) dalle (4.21) si è assunto che la matrice $[I - F]$ sia non singolare. Usando i risultati del precedente probl. 3, mostrare che per un grafo di flusso calcolabile questa assunzione è sempre valida.
5. Nelle eq. (4.29) la rappresentazione matriciale del grafo di flusso è in una forma in cui $w(n)$ è espresso esplicitamente in termini soltanto dei valori passati delle variabili e del valore attuale dell'ingresso. In questo problema si vuole modificare tale rappresentazione in modo tale da ottenerne una in cui il vettore di nodo $w(n_0)$ al tempo n_0 può essere calcolato dal vettore di nodo $w(n_0)$ al tempo n_0 e dal vettore di ingresso $s(n)$ per $n_0 \leq n \leq n_1$.
- (a) Modificare l'eq. (4.29a) per esprimere $w(n)$ in termini di $w(n-2)$ e degli ingressi $x(n-1)$ e $x(n)$.
- (b) Sia $n_1 = n_0 + M$ dove M è costante. Generalizzando la procedura ed il risultato della parte (a), determinare una rappresentazione matriciale basata sulla eq. (4.29a) ma nella quale $w(n_1)$ è espresso in funzione di $w(n_0)$ e degli ingressi $x(n_1 - M)$, $x(n_1 - M + 1)$, ..., $x(n_1)$.

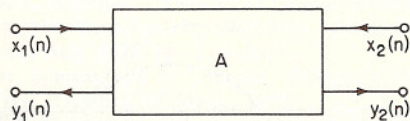


Fig. P4.6-1

6. In figura P4. 6-1 è indicata una rete di elaborazione numerica A con due ingressi, $x_1(n)$ e $x_2(n)$, e due uscite $y_1(n)$ e $y_2(n)$. La rete A può essere descritta con le due equazioni alle porte

$$Y_1(z) = H_1(z)X_1(z) + H_2(z)X_2(z)$$

$$Y_2(z) = H_3(z)X_1(z) + H_4(z)X_2(z)$$

dove

$$H_1(z) = \frac{1}{1 - \frac{1}{2}z^{-1}}$$

$$H_2(z) = 1$$

$$H_3(z) = \frac{1 + 2z^{-1}}{1 + \frac{1}{2}z^{-1}}$$

$$H_4(z) = \frac{1}{1 + \frac{1}{2}z^{-1}}$$

- (a) Disegnare il grafo di flusso che realizza la rete. Il coefficiente di trasmissione di ogni ramo deve essere una costante oppure una costante per z^{-1} . Funzioni di z^{-1} di ordine più elevato non possono essere usate come coefficienti di trasmissione dei rami.

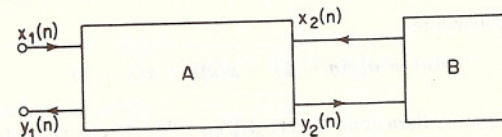


Fig. P4.6-2

- (b) Scrivere il sistema di equazioni corrispondenti alla rete trovata nella parte (a) nella forma delle equazioni (4.21).
- (c) Determinare un ordinamento dei nodi nel grafo di flusso in corrispondenza del quale il grafo di flusso non è calcolabile e determinare un ordinamento dei nodi per il quale il grafo di flusso è calcolabile.
- (d) Si desidera ora connettere una rete B alla rete A come è indicato in fig. P4.6-2. La rete B non ha nodi sorgente interni ed è caratterizzata dalla funzione di trasferimento $X_2(z) = Y_2(z)H_B(z)$. Si osservi che a causa della notazione usata per la rete A , $y_2(n)$ indica l'ingresso alla rete B e $x_2(n)$ è l'uscita. Esaminando il grafo di flusso della parte (a) o le equazioni della parte (b), determinare una condizione necessaria e sufficiente su $H_B(z)$ in modo tale che il sistema complessivo di fig. P4.6-2 sia calcolabile. Se si incontrano difficoltà, provare prima le due possibilità $H_B(z) = 1$ e $H_B(z) = z^{-1}$.
7. Si consideri il sistema lineare causale a tempo discreto descritto dall'equazione alle differenze

$$y(n) - \frac{3}{4}y(n-1) + \frac{1}{8}y(n-2) = x(n) + \frac{1}{3}x(n-1)$$

Disegnare un grafo di flusso di segnale che realizzi questo sistema in ognuna delle seguenti forme:

- (a) Forma diretta I
(b) Forma diretta II
(c) Cascata
(d) Parallela

Per le forme in cascata ed in parallelo usare soltanto blocchi del primo ordine.

8. Per molte applicazioni è utile avere una rete numerica che generi una sequenza sinusoidale. Uno dei modi possibili per far ciò è utilizzare una rete la cui risposta all'impulso è $e^{j\omega_0 n}u(n)$. Le parti reale ed immaginaria di questa risposta sono allora, rispettivamente, $(\cos \omega_0 n)u(n)$ e $(\sin \omega_0 n)u(n)$.

Nel realizzare un sistema con risposta all'impulso complessa, la parte reale e la parte immaginaria sono trattate come uscite separate. Scrivendo prima l'equazione alle differenze complessa che produce la risposta desiderata, ed eguagliando le parti reali e le parti immaginarie dei due membri, disegnare una rete numerica che realizza questo sistema. Tale rete può avere soltanto coefficienti reali. Questa realizzazione è spesso indicata come *oscillatore in forma accoppiata*.

9. Il sistema con funzione di trasferimento $H(z) = (z^{-1} - a)/(1 - az^{-1})$ è un sistema passa-tutto, cioè la sua risposta in frequenza ha modulo unitario.

- (a) Disegnare una rete che realizza questo sistema nella forma diretta II.
(b) Nel realizzare la rete della parte (a), i coefficienti devono, naturalmente, essere soggetti a quantizzazione. Con i coefficienti quantizzati la rete della parte (a) corrisponde ancora ad una rete passa-tutto?

L'equazione alle differenze che lega l'uscita all'ingresso del sistema passa-tutto può essere espressa come

$$y(n) - ay(n-1) = x(n-1) - ax(n)$$

o equivalentemente

$$y(n) = a[y(n-1) - x(n)] + x(n-1) \quad (\text{P4.9-1})$$

- (c) Disegnare una realizzazione di rete dell'eq. (P4.9-1) che richieda due rami con ritardo ma solo un ramo con coefficiente di trasmissione diverso da +1 o da -1.

- (d) Con i coefficienti quantizzati, la rete della parte (c) corrisponde ancora ad una rete passa-tutto?

Il principale svantaggio della realizzazione della parte (c) rispetto a quella della parte (a) è che essa richiede due ritardi. In qualche applicazione, comunque, è necessario realizzare una cascata di blocchi passa-tutto. Per N blocchi passa-tutto in cascata è possibile utilizzare una realizzazione di ciascun blocco nella forma determinata nella parte (c) ma usando soltanto $(N+1)$ rami con ritardo. Questo si ottiene, essenzialmente, mettendo in comune un ritardo fra più blocchi.

- (e) Si consideri il sistema passa-tutto con funzione di trasferimento

$$H(z) = \frac{z^{-1} - a}{1 - az^{-1}} \frac{z^{-1} - b}{1 - bz^{-1}} \quad (\text{P4.9-2})$$

Disegnare una rete che realizza questo sistema connettendo in cascata due reti della forma ottenuta nella parte (c) in modo tale che siano richiesti solo tre rami con ritardo.

- (f) Con i coefficienti quantizzati, la rete della parte (e) corrisponde ancora ad una rete passa-tutto?
10. È stata proposta (S.K. Mitra e R.J. Sherwood, « Canonic Realization of Digital Filters Using the Continued Fraction Expansion », *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, 1972, pp. 185-194) una classe di strutture di filtri numerici basate sulle espansioni di frazioni continue. Esistono molte forme per tali strutture, e in questo problema ne illustreremo una particolare.

Si consideri una funzione di trasferimento $H(z)$ della forma

$$H(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_M z^{-M}}{1 - a_1 z^{-1} - \dots - a_N z^{-N}}$$

dove si assume che $b_0 \neq 0$, $a_N \neq 0$ e $M \leq N$.

Moltiplicando numeratore e denominatore di $H(z)$ per z^N essa può essere espressa come

$$H(z) = \frac{b_0 z^N + b_1 z^{N-1} + \dots + b_M z^{N-M}}{z^N - a_1 z^{N-1} - \dots - a_N}$$

Effettuando la divisione tra numeratore e denominatore si ottiene

$$H(z) = A_0 + G_0(z)$$

dove $A_0 = b_0$ e $G_0(z)$ assume in generale la forma

$$G_0(z) = \frac{c_1 z^{N-1} + \dots + c_M}{z^N - a_1 z^{N-1} - \dots - a_N}$$

Ora se $c_1 \neq 0$ e se si divide il denominatore per il numeratore, si può esprimere $G_0(z)$ come

$$G_0(z) = \frac{1}{A_1 + B_1 z + G_1(z)}$$

dove $G_1(z)$ assume la forma

$$G_1(z) = \frac{d_2 z^{N-2} + \dots + d_N}{c_1 z^{N-1} + \dots + c_M}$$

Si può ripetere la divisione del denominatore per il numeratore ottenendo

$$G_1(z) = \frac{1}{A_2 + B_2 z + G_2(z)}$$

Così, assumendo che le funzioni razionali $\{G_k(z)\}$, $k = 0, 1, \dots, N$, ottenute con questo procedimento abbiano il numeratore di grado $N-k-1$ ed il denominatore di grado $N-k$, allora $H(z)$ può essere espressa come

$$H(z) = A_0 + \frac{1}{A_1 + B_1 z + \frac{1}{A_2 + B_2 z + \frac{1}{\ddots + \frac{1}{A_N + B_N z}}}} \quad (\text{P4.10-1})$$

Per ricavare una rete basata sulla (P4.10-1), basta realizzare soltanto la funzione di trasferimento

$$G_k(z) = \frac{1}{A_{k+1} + B_{k+1} z + G_{k+1}(z)} \quad (\text{P4.10-2})$$

Moltiplicando numeratore e denominatore della (P4.10-2) per $(1/B_{k+1})z^{-1}$ si ottiene

$$G_k(z) = \frac{(1/B_{k+1})z^{-1}}{1 + (A_{k+1}/B_{k+1})z^{-1} + (1/B_{k+1})z^{-1}G_{k+1}(z)} \quad (\text{P4.10-3})$$

Una rete che realizza la (P4.10-3) è mostrata in fig. P4.10

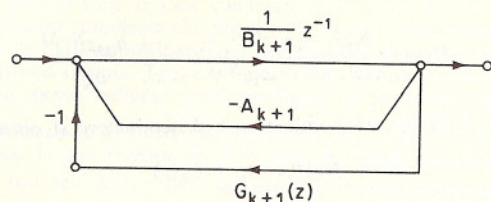


Fig. P4.10

(a) Poiché ogni $G_k(z)$ può essere realizzata da una rete simile, ciò suggerisce una struttura completa per $H(z)$ espressa nella forma (P4.10-1). Assumendo che N sia dispari, disegnare la rete per tale struttura. Ogni ramo in questa rete deve avere un coefficiente di trasmissione che è una costante o una costante moltiplicata per z^{-1} .

(b) Per il sistema del secondo ordine con funzione di trasferimento

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

esprimere $H(z)$ nella forma (P4.10-1).

(c) Disegnare la rete che realizza il sistema della parte (b) nella forma che è stata determinata nella parte (a).

11. Il processo di produzione della voce può essere modellato con un sistema lineare che rappresenta il tratto vocale, eccitato da soffi d'aria emessi attraverso le corde vocali. Per sintetizzare la voce con un calcolatore numerico, un possibile approccio consiste nel rappresentare il tratto vocale come una cascata di tubi acustici cilindrici di uguale lunghezza ma sezione differente, come mostrato in fig. P4.11.

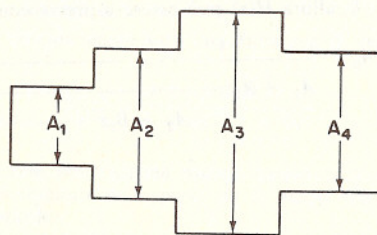
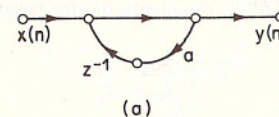


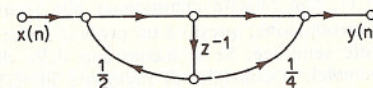
Fig. P4.11

Assumiamo che si voglia simulare questo sistema in termini della portata rappresentante il flusso d'aria. L'ingresso è accoppiato al tratto vocale attraverso una piccola apertura, costituita dalle corde vocali. Assumeremo che l'ingresso sia rappresentato dalla variazione di portata del flusso d'aria al bordo sinistro, ma che la condizione al contorno per le onde viaggianti nel medesimo bordo sia di portata netta nulla. Ciò è analogo ad una linea elettrica di trasmissione pilotata da un generatore di corrente. Come uscita del sistema prendiamo la portata in corrispondenza del bordo destro. Assumiamo infine che ogni sezione sia senza perdite.

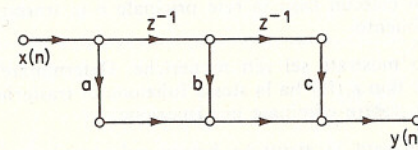
Ad ogni interfaccia tra le diverse sezioni viene trasmessa alla sezione successiva un'onda viaggiante diretta con un certo coefficiente e questa stessa viene



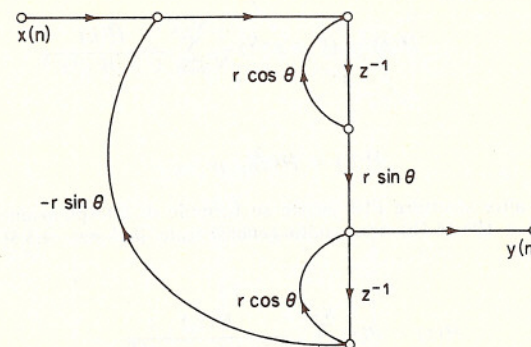
(a)



(b)



(c)



(d)

Fig. P4.12

riflessa come un'onda viaggiante inversa con un diverso coefficiente. In maniera simile, un'onda viaggiante inversa che arriva ad un'interfaccia è trasmessa con un coefficiente e riflessa con un altro. Più precisamente, se consideriamo un'onda viaggiante diretta f_+ in un tubo con sezione A_1 che arriva all'interfaccia con un tubo di sezione A_2 , allora l'onda viaggiante diretta trasmessa è $(1 + \alpha)f_+$ e quella riflessa è αf_+ , dove

$$\alpha = \frac{A_2 - A_1}{A_1 + A_2}$$

Si consideri che la lunghezza di ogni sezione sia 3.4 cm con la velocità del suono nell'aria di 34.000 cm/s. Disegnare una rete numerica che realizza il tubo a quattro sezioni di fig. P4.11, con l'uscita campionata alla frequenza di 20 kHz.

Nonostante la lunga introduzione, questo è un problema che può essere risolto in maniera ragionevolmente semplice. Se si incontrano delle difficoltà nel ragionare in termini di tubi acustici, si consiglia di ragionare in termini di sezioni di linee di trasmissione con impedenze caratteristiche differenti. Proprio come nel caso delle linee di trasmissione, risulta difficile esprimere la risposta all'impulso in forma chiusa. Disegnare direttamente la rete in base a considerazioni fisiche, in termini di impulsi viaggianti in senso diretto e inverso in ogni sezione.

12. In fig. P4.12(a)-(d) sono mostrate alcune reti. Determinare la trasposta di ognuna di esse e verificare che in ciascun caso la rete originale e la trasposta hanno la stessa funzione di trasferimento.
13. In fig. P4.13-1(a)-(f) sono mostrate sei reti numeriche. Determinare quale delle ultime cinque [cioè da (b) fino a (f)] ha la stessa funzione di trasferimento di (a). Alcune possibilità possono essere eliminate per ispezione.
14. Nel par. 4.5.4 è stata derivata la struttura basata sul campionamento in frequenza per filtri FIR, ricavata da un'espansione della funzione di trasferimento $H(z)$ in termini della DFT della risposta all'impulso o equivalentemente in termini di N campioni equispaziati sul circolo unitario. L'espansione usata è quella data nella (4.43), cioè

$$H(z) = (1 - z^{-N}) \frac{1}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - W_N^k z^{-1}} \quad (\text{P4.14-1a})$$

dove

$$\tilde{H}(k) = H(z)|_{z=W_N^{-k}} \quad (\text{P4.14-1b})$$

Per derivare altre strutture FIR basate su formule di interpolazione polinomiale, le (P4.14-1a) e (P4.14-1b) sono state generalizzate (nel par. 4.5.5) nelle (4.53), (4.54) e (4.55):

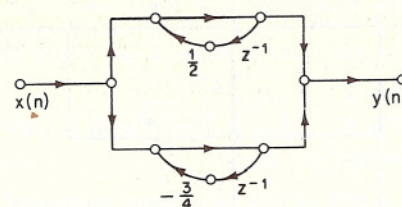
$$H(z) = P(z) \sum_{k=0}^{N-1} \frac{H(z_k)}{P_k(z_k)(1 - z_k z^{-1})} \quad (\text{P4.14-2a})$$

dove

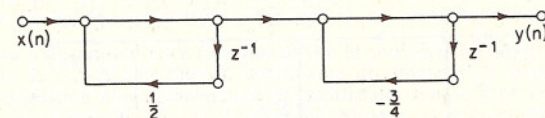
$$P(z) = \prod_{k=0}^{N-1} (1 - z_k z^{-1}) \quad (\text{P4.14-2b})$$

e

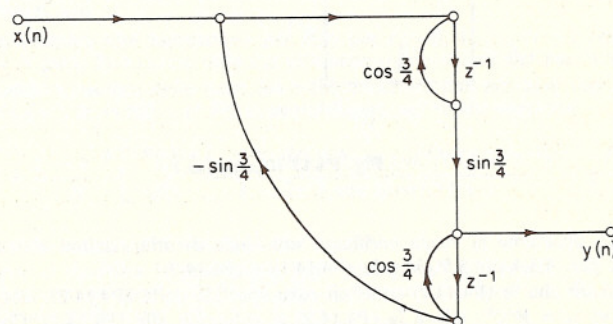
$$P_k(z) = \prod_{\substack{i=0 \\ i \neq k}}^{N-1} (1 - z_i z^{-1}) \quad (\text{P4.14-2c})$$



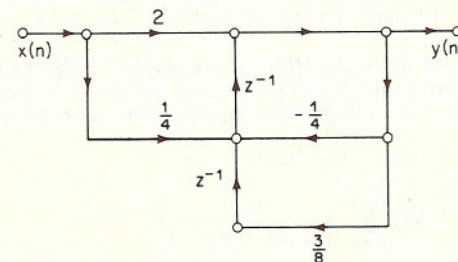
(a)



(b)

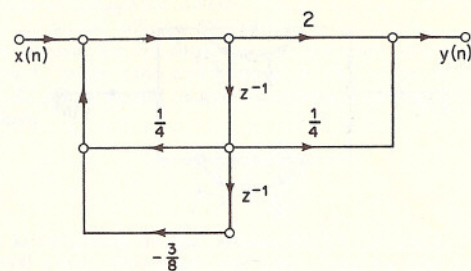


(c)

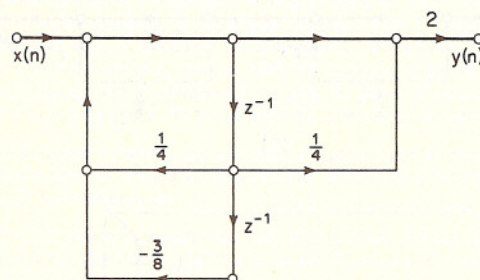


(d)

Fig. P4.13



(e)



(f)

Fig. P4.13 (Continuaz.)

In questo problema si vuole verificare una serie di affermazioni che sono state fatte nei par. 4.5.4 e 4.5.5 circa le (P4.14-1) e (P4.14-2).

- Mostrare che le (P4.14-1) sono un caso speciale delle (P4.14-2), cioè mostrare che se $z_k = W_N^k$, allora le (P4.14-2) si riducono alle (P4.14-1). Come conseguenza, basta concentrarsi solo sul caso più generale rappresentato dalle (P4.14-2).
- Mostrare che $H(z)$ come è espressa dalle (P4.14-2) è un polinomio in z^{-1} di grado $(N-1)$. Per far ciò è necessario mostrare che non ci sono poli eccetto che per $z=0$ e che non ci sono potenze di z^{-1} di ordine più elevato di $(N-1)$.
- Si desidera infine verificare che $H(z)$ come è espressa dalle (P4.14-2) fornisce i valori corretti nei punti di campionamento nel piano z , cioè si desidera verificare che per $z=z_i$ il secondo membro della (P4.14-2a) si riduce a $H(z_i)$.

Mostrare, usando la regola di L'Hopital, che

$$\lim_{z \rightarrow z_i} \left[P(z) \sum_{k=0}^{N-1} \frac{H(z_k)}{P_k(z_k)(1 - z_k z^{-1})} \right] = H(z_i)$$

- Consideriamo un filtro con risposta all'impulso finita e per il quale la risposta in frequenza ha la proprietà che

$$H(e^{j\omega}) = |H(e^{j\omega})| e^{-j\omega n_0}$$

dove n_0 non è necessariamente un intero. Sia N la lunghezza della risposta all'impulso. Ricordiamo che questa è completamente specificata da N campioni di $H(e^{j\omega})$ presi a $\omega = 2\pi k/N$, $k = 0, 1, \dots, N-1$.

- Rappresentare $|H(e^{j\omega})|$ per il caso $N = 15$, $n_0 = 0$, e

$$|H_k| = |H(e^{j(2\pi k/15)})| = \begin{cases} 1, & k = 0 \\ 1/2, & k = 1, 14 \\ 0, & \text{altrove} \end{cases}$$

- Scrivere un'espressione generale per $h(n)$ in termini degli H_k (non si assuma $n_0 = 0$).
 - Rappresentare $h(n)$ per i casi (1) $n_0 = (N-1)/2 = 7$, e (2) $n_0 = N/2 = 15/2$, dove gli $|H_k|$ sono come nella parte (a).
 - Disegnare il grafo di flusso che realizza questo sistema quando $N = 15$, $n_0 = 15/2$, e gli $|H_k|$ sono come in (a). Questa realizzazione deve essere nella forma di un filtro ricorsivo, cioè basato sul campionamento in frequenza.
- Un sistema FIR causale ed a fase lineare ha la proprietà che $h(n) = h(N-1-n)$ per $n = 0, 1, \dots, N-1$. Questa condizione di simmetria è stata usata nel par. 4.5.3 per mostrare che i sistemi che la soddisfano hanno fase lineare corrispondente ad un ritardo di $(N-1)/2$ campioni. Questa condizione dà origine ad una notevole semplificazione della realizzazione basata sul campionamento in frequenza espressa dalle (4.49) e (4.50).
 - Usando la precedente condizione di fase lineare, mostrare che per N pari è $\tilde{H}(N/2) = 0$.
 - Determinare una espressione per $\theta(k)$ per $k = 0, 1, \dots, N-1$ che sia valida per N pari. Può essere utile far riferimento ai risultati del par. 4.5.3.
 - Usando i risultati delle parti (a) e (b), mostrare che per $h(n)$ con fase lineare e N pari, le (4.49) e (4.50) si semplificano nel modo seguente

$$H(z) = \frac{1 - z^{-N}}{N} \left[\sum_{k=1}^{(N/2)-1} \frac{(-1)^k |\tilde{H}(k)| 2 \cos(\pi k/N) (1 - z^{-1})}{1 - 2z^{-1} \cos(2\pi k/N) + z^{-2}} + \frac{\tilde{H}(0)}{1 - z^{-1}} \right]$$

(è stato assunto per comodità che $r = 1$).

- come si modifica la precedente espressione se N è dispari?
 - Disegnare un grafo di flusso che rappresenti la funzione di trasferimento ricavata nella parte (c).
 - Se un numero sufficiente di coefficienti $|\tilde{H}(k)|$ sono nulli, la struttura basata sul campionamento in frequenza può richiedere un minor numero di operazioni aritmetiche che non la realizzazione in forma diretta. Quanti coefficienti possono essere non nulli e dare ancora luogo a un numero di addizioni e moltiplicazioni minore che nella forma diretta?
- Nel par. 4.5.4 e nel precedente probl. 16 è stata ottenuta una classe di strutture basate sul campionamento in frequenza, che si ricavano campionando la risposta in frequenza di un sistema FIR alle frequenze $\omega_k = 2\pi k/N$, $k = 0, 1, \dots, N-1$. Un secondo tipo di struttura basata sul campionamento in frequenza può essere ottenuta campionando la risposta in frequenza di un sistema FIR alle N frequenze

$$\omega_k = \frac{2\pi}{N} \left(k + \frac{1}{2} \right), \quad k = 0, 1, \dots, N-1$$

Definiamo campioni in frequenza di tipo 2 i seguenti

$$\tilde{H}(k) = H(z) \big|_{z=e^{j\omega_k}}, \quad k = 0, 1, \dots, N-1$$

- Esprimere la risposta all'impulso $h(n)$ in termini dei campioni in frequenza del tipo 2.

- (b) Seguendo la stessa procedura adottata nel paragrafo 3.5, mostrare che $H(z)$ può essere espressa come

$$H(z) = \frac{1 + z^{-N}}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - e^{j(\pi/N)} W_N^{-k} z^{-1}}$$

dove $W_N^{-k} = e^{j(2\pi/N)k}$.

- (c) Se si esprimono i campioni in frequenza del tipo 2 in forma polare come

$$\tilde{H}(k) = |\tilde{H}(k)| e^{j\theta(k)}$$

determinare le condizioni su $|\tilde{H}(k)|$ e $\theta(k)$ che assicurano che $h(n)$ è reale.

- (d) Usare le condizioni della parte (c) per mostrare che per N pari è

$$H(z) = \frac{1 + z^{-N}}{N} \left[\sum_{k=0}^{(N/2)-1} \frac{2|\tilde{H}(k)| [\cos(\theta(k)) - z^{-1} \cos(\theta(k) - (2\pi/N)(k + \frac{1}{2}))]}{1 - 2z^{-1} \cos[(2\pi/N)(k + \frac{1}{2})] + z^{-2}} \right]$$

- (e) Come si modifica la precedente espressione se N è dispari?
 (f) Se il sistema ha fase lineare con ritardo di $(N-1)/2$ campioni, scrivere un'espressione di $\theta(k)$ quando N è pari.
 (g) Usare i risultati della parte (f) per mostrare che per N pari e fase lineare è

$$H(z) = \frac{1 + z^{-N}}{N} \left[\sum_{k=0}^{(N/2)-1} \frac{2|\tilde{H}(k)| (-1)^k (1 + z^{-1}) \sin(\pi(k + \frac{1}{2})/N)}{1 - 2z^{-1} \cos(2\pi(k + \frac{1}{2})/N) + z^{-2}} \right]$$

- (h) Come si modifica il risultato della parte (g) se N è dispari?

Questo problema ed il 16 sono basati sul lavoro di L. Rabiner e R. Schaffer, « Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques », *IEEE Trans. Audio Electroacoust.*, Vol. AU-19, No. 3, Sept. 1971.

18. (a) Verificare che il teorema di Tellegen nella forma (4.62) è valido per le due reti di Fig. 4.39.
 (b) Ripetere la parte (a) per il teorema di Tellegen nella forma (4.65).
 19. Nel par. 4.7.3 è stata presentata un'espressione [la (4.87)] che fornisce la variazione nella funzione di trasferimento di un grafo lineare di flusso di segnale causata da una grande variazione nel coefficiente di trasmissione di un ramo. In questo problema si desidera ricavare quel risultato.

Per derivare questa relazione iniziamo con l'esprimere ΔT_{ab} in serie di Taylor come

$$\Delta T_{ab} = \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k T_{ab}}{\partial F_{nm}^k} [\Delta F_{nm}]^k \quad (\text{P4.19-1})$$

dove T_{ab} indica la funzione di trasferimento dal nodo a al nodo b quando il coefficiente di trasmissione del ramo che unisce il nodo n con il nodo m è dato da F_{nm} e ΔT_{ab} indica la variazione nella funzione di trasferimento quando il coefficiente di trasmissione del ramo cambia di ΔF_{nm} .

- (a) Usare la (4.85) per mostrare che $\partial^2 T_{ab} / \partial F_{nm}^2$ può essere espresso come

$$\frac{\partial^2 T_{ab}}{\partial F_{nm}^2} = 2 T_{mn} T_{an} T_{mb}$$

- (b) Mostrare in maniera simile che

$$\frac{\partial^3 T_{ab}}{\partial F_{nm}^3} = 3! T_{mn}^2 T_{an} T_{mb}$$

I risultati delle parti (b) e (c) si generalizzano nella seguente espressione

$$\frac{\partial^k T_{ab}}{\partial F_{nm}^k} = k! T_{mn}^{(k-1)} T_{an} T_{mb}, \quad n \geq 1 \quad (\text{P4.19-2})$$

- (c) Combinando le (P4.19-1) e (P4.19-2), mostrare che è

$$\Delta T_{ab} = \frac{T_{an} T_{mb} \Delta F_{nm}}{1 - T_{mn} \Delta F_{nm}} \quad (\text{P4.19-3})$$

20. Usare la (4.87) per mostrare che per un filtro numerico non ricorsivo, una grande variazione nel coefficiente di trasmissione di un ramo genera una variazione proporzionale nella funzione di trasferimento globale, cioè mostrare che ΔT_{ab} è proporzionale a ΔF_{nm} .

5. TECNICHE DI PROGETTO DI FILTRI NUMERICI

5.0 INTRODUZIONE

Nella sua accezione più generale, un filtro numerico è un sistema a tempo discreto invariante alla traslazione realizzato usando un'aritmetica a precisione finita. Il progetto dei filtri numerici richiede tre passi fondamentali: (1) la specificazione delle proprietà desiderate del sistema; (2) l'approssimazione di tali specifiche per mezzo di un sistema causale a tempo discreto; e (3) la realizzazione del sistema usando l'aritmetica a precisione finita. Benché questi tre passi non siano certamente indipendenti l'uno dall'altro, riteniamo conveniente concentrare la nostra attenzione in questo capitolo principalmente sul secondo passo, in quanto il primo è molto dipendente dalle particolari applicazioni mentre il terzo è trattato nei capitoli 4 e 9.

In pratica si presenta spesso il caso che il filtro numerico desiderato debba essere usato per filtrare un segnale numerico derivato da un segnale analogico attraverso un campionamento periodico. Le specifiche sia dei filtri analogici che numerici sono date spesso (ma non sempre) nel dominio della frequenza, come è il caso, per esempio, dei filtri passa-basso, passa-banda e passa-alto, cioè dei filtri selettivi in frequenza. Una volta data la velocità di campionamento, è immediato passare dalle specifiche in frequenza sul filtro analogico alle specifiche in frequenza sul corrispondente filtro numerico, essendo le frequenze analogiche definite in termini di Hertz e quelle numeriche in termini di angoli lungo la circonferenza unitaria con il punto $z = -1$ corrispondente a metà della frequenza di campionamento. Esistono tuttavia applicazioni nelle quali il segnale numerico da filtrare non è derivato da un segnale analogico per mezzo di un campionamento periodico, e inoltre, come si è osservato nel par. 1.7, esistono diversi altri modi, oltre il campionamento periodico, per rappresentare funzioni del tempo analogiche per mezzo di sequenze. In aggiunta a ciò va detto che per la maggior parte delle tecniche di progetto che discuteremo, il periodo di campionamento non gioca alcun ruolo nel procedimento di approssimazione. Pertanto il punto di vista che genera minor confusione riguardo al progetto dei filtri numerici è quello di considerare i filtri specificati in termini di angoli lungo la circonferenza unitaria piuttosto che in termini di frequenze analogiche.

Un problema a sé è quello di determinare un appropriato insieme di specifiche sul filtro numerico da progettare. Nel caso di un filtro passa-basso, per esempio, le specifiche prendono spesso la forma di un insieme di limiti di tolleranza, come quello rappresentato nella fig. 5.1¹. La curva tratteggiata rappresenta la risposta in frequenza di un sistema che soddisfa le specifiche richieste. In questo caso è assegnata una banda passante all'interno della quale il modulo della risposta deve approssimare 1 con un errore di $\pm \delta_1$, cioè

$$1 - \delta_1 \leq |H(e^{j\omega})| \leq 1 + \delta_1, \quad |\omega| \leq \omega_p$$

È assegnata inoltre una *banda oscura* all'interno della quale il modulo della risposta deve approssimare zero con un errore minore di δ_2 , cioè

$$|H(e^{j\omega})| \leq \delta_2, \quad \omega_s \leq |\omega| \leq \pi$$

La frequenza di taglio per la banda passante ω_p e la frequenza di taglio per la banda oscura ω_s sono date in termini di angolo nel piano z . Affinché

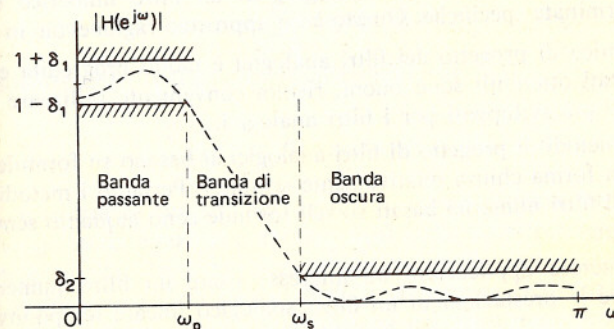


Fig. 5.1 Limiti di tolleranza nell'approssimazione di un filtro ideale passa-basso.

sia possibile approssimare in questo modo il filtro ideale passa-basso, occorre anche assegnare una *banda di transizione* di larghezza non nulla ($\omega_s - \omega_p$) nella quale il modulo della risposta cali gradualmente dalla banda passante alla banda oscura. Molti dei filtri usati in pratica sono specificati proprio da un simile schema di tolleranze, senza nessun vincolo sulla risposta di fase all'infuori di quelli imposti dai requisiti di causalità e stabilità, vale a dire che i poli della funzione di trasferimento devono trovarsi internamente al cerchio unitario.

Dato un insieme di specifiche nella forma rappresentata in fig. 5.1, il passo successivo è quello di trovare un sistema lineare a tempo discreto la cui risposta in frequenza cada all'interno delle tolleranze prescritte. A questo punto il problema del progetto del filtro diventa un problema

¹ In questa figura i limiti degli errori di approssimazione tollerabili sono indicati con le orizzontali tratteggiate. Si noti anche che è sufficiente disegnare le specifiche del filtro solo per $0 \leq \omega \leq \pi$, in quanto il resto può essere dedotto in base alle note proprietà di simmetria.

di approssimazione. Nel caso dei sistemi IIR si tratta di approssimare la risposta in frequenza desiderata mediante una funzione razionale, mentre nel caso FIR si ha a che fare con una approssimazione polinomiale. Per ragioni di convenienza distingueremo nel seguito fra quelle tecniche di progetto che sono adatte ai filtri IIR e quelle adatte ai filtri FIR. Verranno prese in esame numerose tecniche di progetto per entrambi i tipi di filtri. Sono tecniche che vanno dalle procedure in forma chiusa, che richiedono la semplice sostituzione delle specifiche nelle formule di progetto, alle tecniche algoritmiche, dove la soluzione si ottiene mediante procedimenti iterativi.

5.1 PROGETTO DI FILTRI NUMERICI IIR DA FILTRI ANALOGICI

L'approccio tradizionale al progetto di filtri numerici IIR comporta la trasformazione di un filtro analogico in un filtro numerico soddisfacente determinate specifiche. Questo è un approccio ragionevole in quanto:

1. La tecnica di progetto dei filtri analogici è molto progredita e, poiché i risultati ottenibili sono buoni, risulta conveniente utilizzare i procedimenti già sviluppati per i filtri analogici.
2. Molti metodi di progetto di filtri analogici si basano su formule di progetto in forma chiusa relativamente semplici. Pertanto i metodi di progetto di filtri numerici basati su tali formule sono alquanto semplici da applicare.
3. In numerose applicazioni ha interesse usare un filtro numerico per simulare le prestazioni di un filtro analogico lineare tempo-invariante.

Si consideri una funzione di trasferimento analogica,

$$H_a(s) = \frac{\sum_{k=0}^M d_k s^k}{\sum_{k=0}^N c_k s^k} = \frac{Y_a(s)}{X_a(s)} \quad (5.1)$$

dove $x_a(t)$ è l'ingresso, $y_a(t)$ l'uscita e $X_a(s)$ e $Y_a(s)$ sono le loro rispettive trasformate di Laplace. Si assuma che $H_a(s)$ sia stata ottenuta per mezzo di uno dei tipici metodi di approssimazione usati nel progetto dei filtri analogici (alcuni esempi sono discussi nel par. 5.2). L'ingresso e l'uscita di un tale sistema sono legati dall'integrale di convoluzione

$$y_a(t) = \int_{-\infty}^{\infty} x_a(\tau) h_a(t - \tau) d\tau \quad (5.2)$$

dove la risposta all'impulso $h_a(t)$ è la trasformata di Laplace inversa di $H_a(s)$. In alternativa, un sistema analogico che abbia una funzione di tra-

sferimento $H_a(s)$ definita dalla (5.1) può essere descritto dall'equazione differenziale

$$\sum_{k=0}^N c_k \frac{d^k y_a(t)}{dt^k} = \sum_{k=0}^M d_k \frac{d^k x_a(t)}{dt^k} \quad (5.3)$$

La corrispondente funzione di trasferimento razionale per filtri numerici ha la forma

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{\sum_{k=0}^N a_k z^{-k}} = \frac{Y(z)}{X(z)} \quad (5.4)$$

L'uscita del sistema è legata all'ingresso dalla somma di convoluzione

$$y(n) = \sum_{k=-\infty}^{\infty} x(k) h(n - k) \quad (5.5)$$

o, equivalentemente, dall'equazione alle differenze

$$\sum_{k=0}^N a_k y(n - k) = \sum_{k=0}^M b_k x(n - k) \quad (5.6)$$

Nel trasformare un sistema analogico in un sistema numerico noi dobbiamo pertanto ottenere o $H(z)$ o $h(n)$ dalle corrispondenti funzioni del filtro analogico. In simili trasformazioni quello che generalmente si richiede è che le proprietà essenziali della risposta in frequenza analogica siano conservate nella risposta in frequenza del filtro numerico risultante. Ciò implica innanzitutto che l'asse immaginario del piano s si mappi nel circolo unitario del piano z . Una seconda condizione è che un filtro analogico stabile dia luogo a un filtro numerico stabile. In altri termini, se il filtro analogico ha poli solo nella metà sinistra del piano s , il filtro numerico dovrà avere poli soltanto internamente al cerchio unitario. Questi vincoli sono fondamentali per tutte le tecniche che discuteremo in questo paragrafo.

5.1.1 Invarianza all'impulso

Un primo procedimento per trasformare un progetto di filtro analogico in un progetto di filtro numerico consiste nello scegliere la risposta all'impulso di quest'ultimo come una sequenza costituita da campioni ugualmente spaziatati della risposta all'impulso del filtro analogico [1 - 4]. Ovvero

$$h(n) = h_a(nT)$$

dove T è il periodo di campionamento.

Si può dimostrare, generalizzando la (1.29), che la trasformata z di $h(n)$ è legata alla trasformata di Laplace di $h_a(t)$ dall'equazione

$$H(z)|_{z=e^{sT}} = \frac{1}{T} \sum_{k=-\infty}^{\infty} H_a\left(s + j\frac{2\pi}{T}k\right) \quad (5.7)$$

Dalla relazione $z = e^{sT}$ si deduce che le strisce di ampiezza $2\pi/T$ nel piano s si mappano nell'intero piano z come mostrato nella fig. 5.2. La metà sinistra di ogni striscia nel piano s viene a mapparsi nella parte interna alla circonferenza unitaria, la metà destra di ogni striscia nel piano s si mappa invece nella parte esterna alla circonferenza unitaria, e l'asse immaginario del piano s si mappa nella circonferenza unitaria in modo tale che ogni segmento di lunghezza $2\pi/T$ si mappa una sola volta lungo tale circonferenza. Dalla (5.7) risulta chiaro che tutte le strisce del

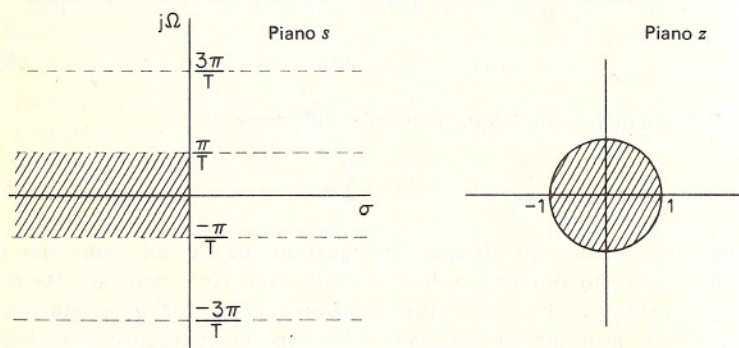


Fig. 5.2 Rappresentazione degli effetti del campionamento periodico.

piano s vanno a sovrapporsi nel piano z per dar luogo, partendo dalla funzione di trasferimento analogica, alla funzione di trasferimento numerica. Pertanto il metodo dell'invarianza all'impulso *non* corrisponde a un semplice mappaggio algebrico del piano s nel piano z .

La risposta in frequenza del filtro numerico è legata alla risposta in frequenza del filtro analogico dalla

$$H(e^{j\omega}) = \frac{1}{T} \sum_{k=-\infty}^{\infty} H_a\left(j\frac{\omega}{T} + j\frac{2\pi}{T}k\right) \quad (5.8)$$

Risulta chiaro, dalla discussione fatta a suo tempo (v. par. 1.7) sul teorema del campionamento, che, se e solo se

$$H_a(j\Omega) = 0, \quad |\Omega| \geq \pi/T$$

allora

$$H(e^{j\omega}) = \frac{1}{T} H_a\left(j\frac{\omega}{T}\right), \quad |\omega| \leq \pi$$

Sfortunatamente qualsiasi filtro analogico che si usi in pratica non sarà limitato in banda, e di conseguenza si ha interferenza tra i vari termini della sommatoria di (5.8), come illustrato nella fig. 5.3.

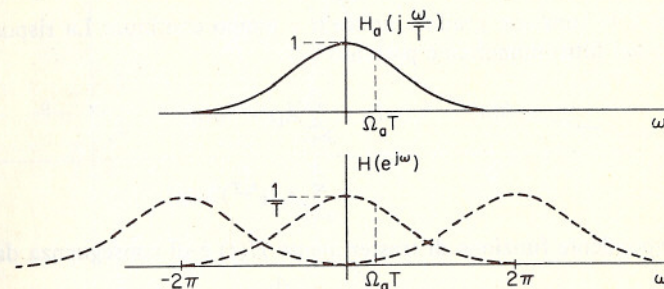


Fig. 5.3 Rappresentazione grafica degli effetti di *aliasing* nella tecnica di progetto dell'invarianza all'impulso

A causa dell'*aliasing* cui dà luogo il processo di campionamento, la risposta in frequenza del filtro numerico risultante non sarà identica alla risposta in frequenza analogica da cui si è partiti. È importante osservare che, se le specifiche del filtro sono date in termini di specifiche sul filtro numerico da progettare, allora un cambiamento nel valore di T non ha effetto sulla quantità di *aliasing* comportata dalla tecnica di progetto dell'invarianza all'impulso. Per esempio, riferendoci alla fig. 5.3, supponiamo che la frequenza di taglio del filtro numerico sia stata scelta pari ad $\Omega_a T$. Questo valore (cioè questa frazione di 2π) è pertanto vincolato ad essere la frequenza di taglio del filtro numerico passa-basso, e, se T viene ridotto, Ω_a deve essere corrispondentemente aumentata nel filtro analogico in modo tale che $\Omega_a T$ rimanga costante ed uguale alla frequenza di taglio specificata per il filtro numerico stesso. Quindi, se T viene ridotto nel tentativo di ridurre l'effetto di *aliasing*, Ω_a deve essere corrispondentemente resa più grande. Se ne deduce che se il filtro numerico da progettare è specificato in termini di frequenze sulla circonferenza unitaria, il parametro T risulta irrilevante nell'ambito della tecnica di progetto dell'invarianza all'impulso e potrebbe pertanto anche porsi uguale ad uno. Poiché tuttavia nell'uso pratico di tale tecnica si include anche T nei parametri del progetto, è importante aver ben presente che tale parametro vi gioca un ruolo del tutto secondario.

Per approfondire l'interpretazione di questa tecnica di progetto in termini di relazione tra i piani s e z , assumiamo che la funzione di trasferimento del filtro analogico sia espressa come sommatoria di fratti semplici, ovvero

$$H_a(s) = \sum_{k=1}^N \frac{A_k}{s - s_k} \quad (5.9)$$

La corrispondente risposta all'impulso è

$$h_a(t) = \sum_{k=1}^N A_k e^{s_k t} u(t)$$

dove $u(t)$ è la funzione gradino unitario a tempo continuo. La risposta all'impulso del filtro numerico è pertanto

$$\begin{aligned} h(n) &= h_a(nT) = \sum_{k=1}^N A_k e^{s_k nT} u(n) \\ &= \sum_{k=1}^N A_k (e^{s_k T})^n u(n) \end{aligned}$$

La corrispondente funzione di trasferimento $H(z)$ è di conseguenza data da

$$H(z) = \sum_{k=1}^N \frac{A_k}{1 - e^{s_k T} z^{-1}} \quad (5.10)$$

Confrontando le espressioni (5.9) e (5.10) osserviamo che un polo in $s = s_k$ nel piano s dà luogo a un polo in $e^{s_k T}$ nel piano z , e che i coefficienti nelle espansioni in fratti semplici di $H_a(s)$ e $H(z)$ sono uguali². Se il filtro analogico è stabile, se cioè la parte reale di s_k è negativa, allora il modulo di $e^{s_k T}$ sarà minore di uno, così che il corrispondente polo del filtro numerico è all'interno del cerchio unitario, e di conseguenza anche il filtro numerico è stabile. Benché i poli nel piano s si mappino in poli nel piano z secondo la relazione $z_k = e^{s_k T}$, è importante riconoscere che il progetto col metodo dell'invarianza all'impulso non corrisponde a un mappaggio dal piano s al piano z secondo quella stessa relazione, né, in effetti, secondo alcun'altra relazione. In particolare, gli zeri della funzione di trasferimento numerica sono una funzione dei poli e dei coefficienti A_k della espansione in fratti semplici e in generale non si mappano da un piano all'altro allo stesso modo dei poli.

ESEMPIO. Come esempio di determinazione di un filtro numerico da un filtro analogico per mezzo dell'invarianza all'impulso, si consideri la funzione di trasferimento analogica $H_a(s)$ data da

$$\begin{aligned} H_a(s) &= \frac{s + a}{(s + a)^2 + b^2} \\ &= \frac{\frac{1}{2}}{s + a + jb} + \frac{\frac{1}{2}}{s + a - jb} \end{aligned}$$

La corrispondente funzione di trasferimento del filtro numerico invariante all'impulso è

$$\begin{aligned} H(z) &= \frac{\frac{1}{2}}{1 - e^{-aT} e^{-jbT} z^{-1}} + \frac{\frac{1}{2}}{1 - e^{-aT} e^{jbT} z^{-1}} \\ &= \frac{1 - (e^{-aT} \cos bT) z^{-1}}{(1 - e^{-aT} e^{-jbT} z^{-1})(1 - e^{-aT} e^{jbT} z^{-1})} \end{aligned}$$

Il filtro numerico ha di conseguenza uno zero nell'origine e uno zero in $z = e^{-aT} \cos bT$.

² Si veda il prob. 5 di questo capitolo per le modifiche da effettuare quando si abbia a che fare con poli di ordine multiplo.

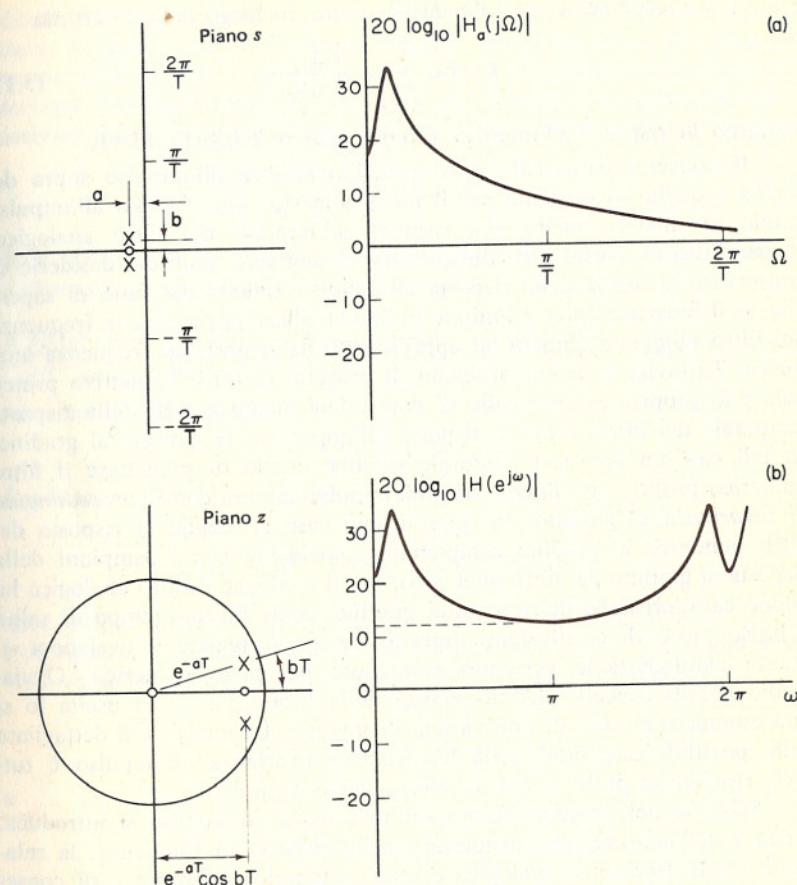


Fig. 5.4 (a) Diagramma di poli e zeri e risposta in frequenza di un sistema analogico del secondo ordine; (b) diagramma di poli e zeri e risposta in frequenza del sistema a tempo discreto ottenuto campionando la risposta all'impulso del sistema precedente.

La fig. 5.4 mostra i diagrammi di poli e zeri di $H_a(s)$ nel piano s e di $H(z)$ nel piano z insieme alle corrispondenti risposte in frequenza analogica e numerica. In questo caso la risposta in frequenza del sistema analogico decade piuttosto lentamente rispetto alla frequenza di campionamento e pertanto gli effetti di aliasing risultano evidenti nella risposta in frequenza numerica.

Va sottolineato che quando il filtro analogico è « sufficientemente limitato in banda », il procedimento sopra illustrato dà luogo a un filtro numerico la cui risposta in frequenza è, in accordo con la (5.8),

$$H(e^{j\omega}) \approx \frac{1}{T} H_a\left(j \frac{\omega}{T}\right)$$

Ne segue che, per elevate frequenze di campionamento (T piccolo), il filtro numerico può avere un guadagno estremamente alto. Per questo

motivo si raccomanda generalmente di usare, in luogo della (5.10), la

$$H(z) = \sum_{k=1}^N \frac{TA_k}{1 - e^{s_k T} z^{-1}} \quad (5.11)$$

Pertanto la risposta all'impulso diventa $h(n) = T h_a(nT)$ [3,4].

Il concetto base della tecnica dell'invarianza all'impulso sopra descritta è quello di scegliere per il filtro numerico una risposta all'impulso simile in qualche modo alla risposta all'impulso del filtro analogico. Spesso l'uso di questo procedimento non è motivato tanto dal desiderio di conservare la forma della risposta all'impulso, quanto dal fatto di sapere che, se il filtro analogico è limitato in banda, allora la risposta in frequenza del filtro numerico tenderà ad approssimare la risposta in frequenza analogica. Tuttavia, in taluni problemi di progetto di filtri, l'obiettivo principale può proprio essere quello di controllare alcuni aspetti della risposta temporale del filtro, come la risposta all'impulso o la risposta al gradino. In tali casi un approccio naturale sarebbe quello di progettare il filtro numerico proprio con l'invarianza all'impulso oppure con il *procedimento di invarianza al gradino*. In quest'ultimo caso si sceglie la risposta del filtro numerico al gradino campionato coincidente con i campioni della risposta al gradino del filtro analogico. In tal modo, se il filtro analogico ha buone caratteristiche di risposta al gradino, come piccolo tempo di salita e basso picco di oscillazione sopra il valore di regime (« overshoot »), queste caratteristiche verranno mantenute nel filtro numerico. Ovviamente questo concetto dell'invarianza della forma d'onda in uscita lo si può estendere al caso di una varietà di ingressi. Un'analisi più dettagliata delle possibili estensioni della tecnica dell'invarianza all'impulso è tuttavia rinviata ai probl. 3 e 4 di questo stesso capitolo.

Sebbene nel progetto basato sull'invarianza all'impulso si introduca, a causa dell'*aliasing*, una distorsione nella risposta in frequenza, la relazione tra la frequenza analogica e quella numerica è lineare e di conseguenza, *aliasing* a parte, la forma della risposta in frequenza si conserva. Ciò non si verifica nei procedimenti che prenderemo fra poco in esame, i quali corrispondono all'uso di trasformazioni algebriche. Va notato in conclusione che la tecnica dell'invarianza all'impulso è chiaramente adatta soltanto nei casi di filtri essenzialmente limitati in banda. Per esempio, filtri del tipo passa-alto o elimina-banda richiederebbero ulteriori limitazioni di banda onde evitare severe distorsioni dovute all'*aliasing*.

5.1.2 Progetti basati sulla soluzione numerica dell'equazione differenziale

Un secondo metodo per ottenere un filtro numerico consiste nell'approssimare le derivate nell'equazione (5.3) con differenze finite. Questa è una procedura standard in analisi numerica [5] e nelle simulazioni di sistemi analogici. Tale procedura può essere motivata dalla nozione intuitiva che la derivata di una funzione analogica del tempo può essere approssi-

simata con la differenza tra valori consecutivi della funzione da differenziare. È lecito attendersi che al crescere della frequenza di campionamento, cioè quando i campioni sono più vicini, l'approssimazione della derivata divenga sempre più accurata. Ad esempio, si supponga che la derivata prima sia approssimata dalla differenza prima « all'indietro » [5]

$$\left. \frac{dy_a(t)}{dt} \right|_{t=nT} \rightarrow \nabla^{(1)}[y(n)] = \frac{y(n) - y(n-1)}{T} \quad (5.12)$$

dove $y(n) = y_a(nT)$. Approssimazioni delle derivate di ordine superiore si ottengono dall'applicazione ripetuta della (5.12), cioè

$$\left. \frac{d^k y_a(t)}{dt^k} \right|_{t=nT} = \frac{d}{dt} \left(\frac{d^{k-1}}{dt^{k-1}} y_a(t) \right) \Big|_{t=nT} \rightarrow \nabla^{(k)}[y(n)] = \nabla^{(1)}[\nabla^{(k-1)}[y(n)]] \quad (5.13)$$

Per comodità definiamo

$$\nabla^{(0)}[y(n)] = y(n) \quad (5.14)$$

Applicando le (5.12) – (5.14) alla (5.3) otteniamo

$$\sum_{k=0}^N c_k \nabla^{(k)}[y(n)] = \sum_{k=0}^M d_k \nabla^{(k)}[x(n)]$$

dove $y(n) = y_a(nT)$ e $x(n) = x_a(nT)$. Notiamo che l'operazione $\nabla^{(1)} []$ è un operatore lineare invariante alla traslazione e che $\nabla^{(k)} []$ può essere visto come la cascata di (k) operatori $\nabla^{(1)} []$. In particolare, si ha

$$\mathfrak{Z}[\nabla^{(1)}[x(n)]] = \left[\frac{1 - z^{-1}}{T} \right] X(z)$$

e

$$\mathfrak{Z}[\nabla^{(k)}[x(n)]] = \left[\frac{1 - z^{-1}}{T} \right]^k X(z)$$

Perciò, effettuando la trasformata z di entrambi i membri otteniamo

$$H(z) = \frac{\sum_{k=0}^M d_k \left[\frac{1 - z^{-1}}{T} \right]^k}{\sum_{k=0}^N c_k \left[\frac{1 - z^{-1}}{T} \right]^k} \quad (5.15)$$

Confrontando la (5.15) con la (5.1), osserviamo che la funzione di trasferimento numerica può essere ottenuta direttamente dalla funzione di trasferimento analogica per mezzo di un cambiamento di variabili

$$s = \frac{1 - z^{-1}}{T} \quad (5.16)$$

così che il procedimento di sostituire le derivate con differenze corrisponde effettivamente ad un mappaggio del piano s nel piano z , secondo la (5.16). Abbiamo indicato in precedenza l'opportunità che l'asse immaginario nel

piano s si mappi sulla circonferenza unitaria nel piano z e che filtri analogici stabili diano luogo a filtri numerici stabili. Per analizzare a questo riguardo il comportamento della trasformazione (5.16) dobbiamo esprimere z in funzione di s , ottenendo

$$z = \frac{1}{1 - sT}$$

Sostituendo $s = j\Omega$, si ha

$$z = \frac{1}{1 - j\Omega T} \quad (5.17)$$

Chiaramente, il luogo dell'asse $j\Omega$ del piano s non è la circonferenza unitaria del piano z , poiché è $|z| \neq 1$ per tutti i valori di Ω nella (5.17). Infatti, possiamo scrivere la (5.17) come

$$\begin{aligned} z &= \frac{1}{2} \left[1 + \frac{1 + j\Omega T}{1 - j\Omega T} \right] \\ &= \frac{1}{2} [1 + e^{j2 \tan^{-1}(\Omega T)}] \end{aligned} \quad (5.18)$$

che corrisponde ad una circonferenza il cui centro è in $z = 1/2$ e il cui raggio è $1/2$, come mostrato in fig. 5.5. Si verifica facilmente che il semi-

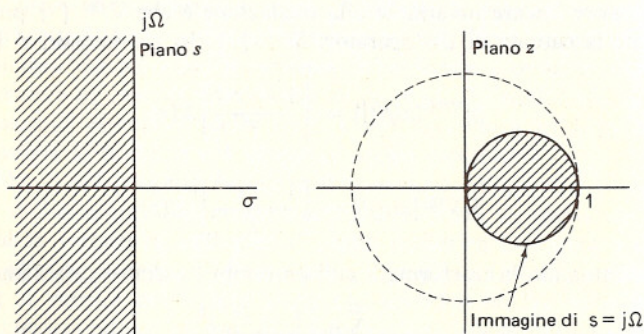


Fig. 5.5 Mappaggio del piano s nel piano z corrispondente all'approssimazione della derivata con la differenza prima «all'indietro».

piano sinistro del piano s si mappa nell'interno della circonferenza piccola e che il semipiano destro del piano s si mappa nell'esterno della stessa circonferenza. Perciò, sebbene questa trasformazione non soddisfi la proprietà che l'asse $j\Omega$ si mappi sulla circonferenza unitaria, essa soddisfa tuttavia la condizione di stabilità, dal momento che i poli nel semipiano sinistro del piano s si mappano all'interno della circonferenza piccola, che è all'interno della circonferenza unitaria.

Vale la pena di mettere in relazione questo risultato con una comune nozione intuitiva. Generalmente si assume che la simulazione su un cal-

colatore numerico dell'elaborazione di un segnale a tempo continuo rappresentato da un'equazione differenziale possa essere effettuata sostituendo le derivate con differenze, se il segnale continuo è campionato con una frequenza sufficientemente elevata. Ad esempio, se desideriamo differenziare un segnale continuo, ci aspettiamo intuitivamente che una approssimazione della derivata possa essere effettuata campionando la funzione a tempo continuo con una spaziatura sufficientemente piccola tra i campioni e formando la differenza prima della sequenza risultante. Per mostrare che effettivamente tale intuizione è in accordo con i risultati ora ottenuti, notiamo innanzitutto che se un segnale analogico a banda limitata è campionato alla frequenza di Nyquist, allora lo spettro è non nullo su tutta la circonferenza unitaria. Al crescere della frequenza di campionamento a partire da quella di Nyquist, cioè al diminuire del periodo di campionamento, la parte non nulla dello spettro del segnale numerico è confinata in una regione sempre più piccola della circonferenza unitaria e, in particolare, se scegliamo un periodo di campionamento sufficientemente piccolo, possiamo concentrare la parte non nulla dello spettro in un intorno di $z = 1$ nel piano z . Corrispondentemente, se T è sufficientemente piccolo nelle formule viste sopra, allora la risposta in frequenza del filtro numerico sarà concentrata sulla circonferenza piccola di fig. 5.5 in un intorno di $z = 1$. Questo è, naturalmente, il punto in cui la circonferenza piccola e la circonferenza unitaria sono tangenti, e, se sia la risposta del filtro che lo spettro del segnale sono concentrati in quella regione, possiamo attenderci che il filtro numerico approssimi accuratamente il filtro analogico.

Nel procedimento precedente, le derivate sono state sostituite con differenze «all'indietro». Una approssimazione alternativa della derivata è la differenza «in avanti». La differenza prima «in avanti» è definita come

$$\Delta^{(1)}[y(n)] = y(n+1) - y(n)$$

Il mappaggio corrispondente a tale approssimazione è esaminato nel probl. 2 di questo capitolo, in cui è mostrato che da questa approssimazione possono risultare filtri numerici instabili.

L'aspetto principale nell'esempio precedente ed anche nel probl. 2 di questo capitolo è che, in contrasto con la tecnica dell'invarianza all'impulso, la riduzione del periodo di campionamento dà luogo teoricamente ad un filtro migliore, dal momento che lo spettro tende ad essere concentrato in una piccola regione della circonferenza unitaria. In generale, tuttavia, l'uso di differenze «in avanti» o «all'indietro» nell'elaborazione numerica dei segnali non è molto raccomandabile, poiché le alte frequenze di campionamento richieste portano ad una rappresentazione molto inefficiente del filtro e del segnale di ingresso. Inoltre, è chiaro che queste procedure sono assai poco soddisfacenti per tutti i filtri ad eccezione dei filtri passa-basso. Perciò, sempre allo scopo di evitare i problemi di *aliasing* propri del metodo dell'invarianza all'impulso, siamo portati a considerare altre trasformazioni.

5.1.3 Trasformazione bilineare

Nel paragrafo precedente si è ricavato un filtro numerico approssimando le derivate con differenze. Un procedimento alternativo è basato sull'integrazione dell'equazione differenziale e sull'uso di un'approssimazione numerica dell'integrale. Ad esempio, si consideri l'equazione del primo ordine

$$c_1 y'_a(t) + c_0 y_a(t) = d_0 x(t) \quad (5.19)$$

dove $y'_a(t)$ è la derivata prima di $y_a(t)$. La corrispondente funzione di trasferimento del sistema analogico è

$$H_a(s) = \frac{d_0}{c_1 s + c_0}$$

Possiamo scrivere $y_a(t)$ come integrale di $y'_a(t)$, nella forma

$$y_a(t) = \int_{t_0}^t y'_a(t) dt + y_a(t_0)$$

In particolare, per $t = nT$ e $t_0 = (n-1)T$, risulta

$$y_a(nT) = \int_{(n-1)T}^{nT} y'_a(\tau) d\tau + y_a((n-1)T)$$

Se l'integrale è approssimato col metodo dei trapezi [5], possiamo scrivere

$$y_a(nT) = y_a((n-1)T) + \frac{T}{2} [y'_a(nT) + y'_a((n-1)T)] \quad (5.20)$$

Poiché dalla (5.19) si ha

$$y'_a(nT) = \frac{-c_0}{c_1} y_a(nT) + \frac{d_0}{c_1} x_a(nT)$$

la (5.20) si può riscrivere

$$[y(n) - y(n-1)] = \frac{T}{2} \left[\frac{-c_0}{c_1} (y(n) + y(n-1)) + \frac{d_0}{c_1} (x(n) + x(n-1)) \right]$$

dove $y(n) = y_a(nT)$ e $x(n) = x_a(nT)$. Prendendo la trasformata z e ricavando $H(z)$ si ha

$$\Rightarrow H(z) = \frac{Y(z)}{X(z)} = \frac{d_0}{c_1 \frac{2}{T} \frac{1-z^{-1}}{1+z^{-1}} + c_0} \quad (5.21)$$

Dalla (5.21) è chiaro che $H(z)$ è ottenuta da $H_a(s)$ con la sostituzione

$$s = \frac{2}{T} \frac{1-z^{-1}}{1+z^{-1}} \quad (5.22)$$

Si può quindi scrivere

$$H(z) = H_a(s) \Big|_{s=\frac{2}{T} \frac{1-z^{-1}}{1+z^{-1}}} \quad (5.23)$$

Si può mostrare che questo vale in generale, poiché un'equazione differenziale di ordine N della forma (5.3) può essere scritta come un sistema di N equazioni del primo ordine della forma (5.19). Ricavando z dalla (5.22) si ha

$$z = \frac{1 + (T/2)s}{1 - (T/2)s} \quad (5.24)$$

Si riconosce che la trasformazione invertibile (5.22) è una trasformazione bilineare [1-4,6]. Per dimostrare che questa trasformazione ha la proprietà di mappare l'asse immaginario del piano s sulla circonferenza unitaria, poniamo $z = e^{j\omega}$. Allora in base alla (5.22) s è dato da

$$\begin{aligned} s &= \frac{2}{T} \frac{1 - e^{-j\omega}}{1 + e^{-j\omega}} \\ &= \frac{2}{T} \frac{j \sin(\omega/2)}{\cos(\omega/2)} \\ &= \frac{2}{T} j \tan(\omega/2) \\ &= \sigma + j\Omega \end{aligned}$$

Perciò, per z sulla circonferenza unitaria, risulta $\sigma = 0$ mentre Ω ed ω sono legati da

$$\frac{T\Omega}{2} = \tan(\omega/2)$$

Questa relazione è riportata nel grafico di fig. 5.6. Dalla figura è chiaro che

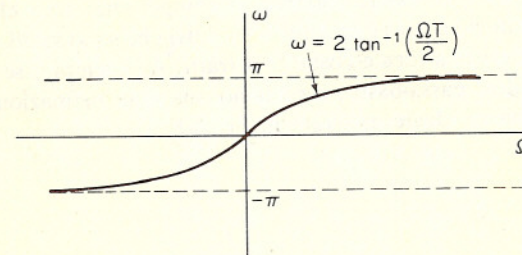


Fig. 5.6 Mappaggio dell'asse delle frequenze analogiche sulla circonferenza unitaria usando la trasformazione bilineare.

l'asse immaginario positivo e quello negativo del piano s sono mappati, rispettivamente, nella metà superiore e in quella inferiore della circonferenza unitaria nel piano z .

In aggiunta al fatto che l'asse immaginario del piano s si mappa sulla circonferenza unitaria nel piano z , il semipiano sinistro del piano s si mappa nell'interno del cerchio unitario e il semipiano destro del piano s si mappa all'esterno, come mostrato in fig. 5.7.

Ciò può essere visto riferendosi alla (5.24). Quando la parte reale di s è negativa, il modulo del rapporto $(1+sT/2)/(1-sT/2)$ è minore di 1, e corrisponde all'interno della circonferenza unitaria. Al contrario, quando la parte reale di s è positiva, il modulo di tale rapporto è maggiore di 1,

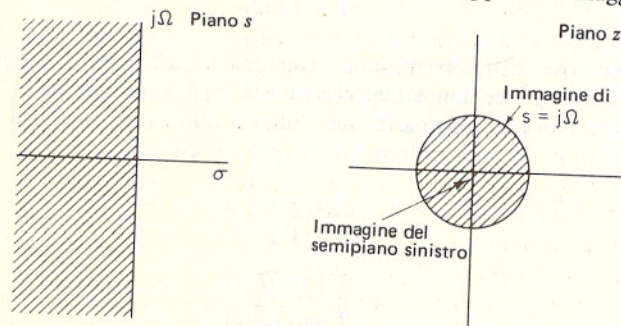


Fig. 5.7 Mappaggio del piano s nel piano z usando la trasformazione bilineare.

e corrisponde all'esterno della circonferenza unitaria. Perciò vediamo che l'uso della trasformazione bilineare fornisce filtri numerici stabili partendo da filtri analogici stabili. Inoltre la trasformazione bilineare evita il problema dell'*aliasing* incontrato con l'uso dell'invarianza all'impulso, poiché mappa l'intero asse immaginario del piano s sulla circonferenza unitaria nel piano z . Il prezzo pagato per questo, tuttavia, è l'introduzione di una distorsione nell'asse frequenza. Di conseguenza, il progetto di filtri numerici per mezzo della trasformazione bilineare è utile solamente quando questa distorsione può essere tollerata o compensata. Una classe particolare di filtri in cui ciò è vero è costituita dai filtri che sono scelti per approssimare una caratteristica ideale costante a tratti. Ad esempio, se desideriamo progettare un filtro passa-basso, cerchiamo una approssimazione alla caratteristica passa-basso ideale mostrata in fig. 5.8.

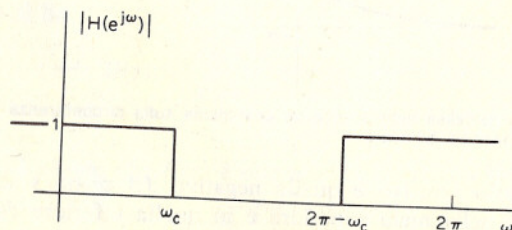


Fig. 5.8 Risposta in frequenza di un filtro passa-basso ideale.

Se fossimo capaci di progettare un filtro passa-basso ideale nel piano s con una frequenza di taglio $\Omega_c = (2/T) \tan(\omega_c/2)$ allora, quando tale progetto fosse mappato nel piano z per mezzo della trasformazione bilineare, ne risulterebbe la caratteristica ideale di fig. 5.8. Naturalmente, non siamo capaci di realizzare un filtro di questo tipo né nel caso analogico né in quello numerico. In generale, dobbiamo approssimare tale caratteristica del filtro, permettendo un certo scostamento rispetto a uno nella banda passante, e rispetto a zero nella banda oscura, con una banda di transizione di larghezza non nulla. La fig. 5.9 rappresenta la trasformazione di una risposta in frequenza analogica e del suo schema di tolleranze nei loro corrispettivi numerici. Se le frequenze critiche del filtro analogico sono pre-distorte come mostrato, allora, quando il filtro analogico è trasformato nel filtro numerico usando la (5.23), quest'ultimo soddisferà le specifiche desiderate.

Tipici progetti di filtri analogici selettivi sono i filtri di Butterworth, di Chebyshev ed ellittici [7,8]. Come vedremo nel prossimo paragrafo, questi metodi di approssimazione analogica hanno formule di progetto in forma chiusa, che rendono il progetto stesso alquanto rapido. Un filtro analogico di Butterworth è monotono nella banda passante e nella banda oscura.

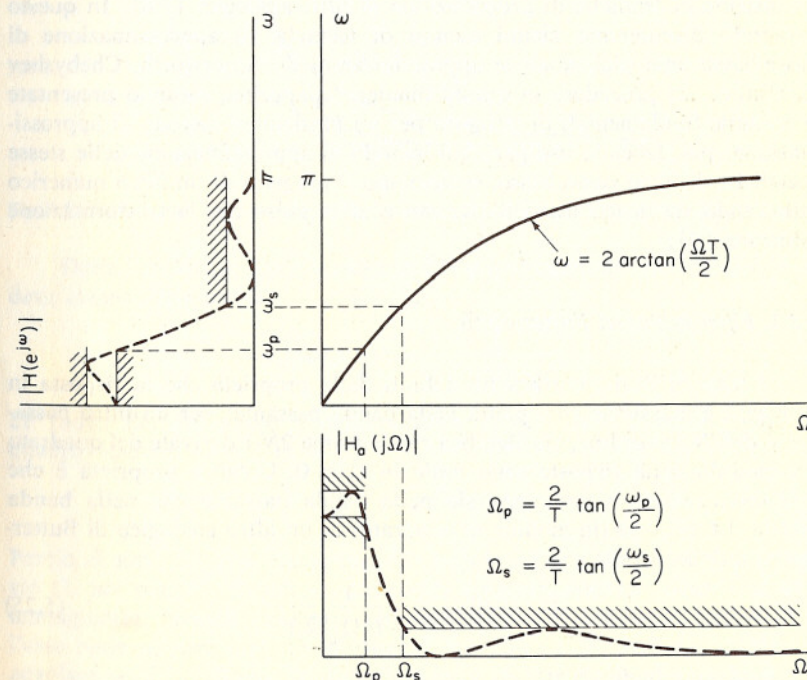


Fig. 5.9 Distorsione di frequenza incontrata nella trasformazione di un filtro passa-basso analogico in un filtro passa-basso numerico. Per ottenere la frequenza di taglio numerica desiderata, le frequenze di taglio analogiche devono essere pre-distorte come indicato.

Un filtro di Chebyshev ha una caratteristica ad oscillazione uniforme nella banda passante e monotona nella banda oscura. Un filtro ellittico è ad oscillazione uniforme sia nella banda passante che nella banda oscura. Chiaramente, queste proprietà saranno conservate quando il filtro è mappato in un filtro numerico mediante la trasformazione bilineare. Ciò è illustrato dalle curve di risposta in frequenza tratteggiate in fig. 5.9.

Sebbene la trasformazione bilineare possa essere usata efficacemente per mappare una caratteristica di ampiezza costante a tratti dal piano s al piano z , la distorsione nell'asse frequenza si manifesterà in termini di distorsione della caratteristica di fase associata col filtro. Se, ad esempio, fossimo interessati ad un filtro passa-basso numerico con una caratteristica di fase lineare, non potremmo ottenere un tale filtro applicando la trasformazione bilineare ad un filtro passa-basso analogico con una caratteristica di fase lineare.

5.2 ESEMPI DI PROGETTO: TRASFORMAZIONE ANALOGICO-NUMERICA

I metodi del paragrafo precedente si basano sulla disponibilità di tutto un insieme di tecniche di progettazione di filtri analogici [7,8]. In questo paragrafo discuteremo alcuni esempi di tecniche di approssimazione di passa-basso analogici, quali le approssimazioni di Butterworth, Chebyshev ed ellittica. Si procederà in questa maniera: dapprima saranno presentate le formule fondamentali di progetto per un particolare metodo di approssimazione; poi, facendo uso per ogni metodo di approssimazione delle stesse specifiche di filtro passa-basso, svolgeremo il progetto di un filtro numerico utilizzando sia la tecnica dell'invarianza all'impulso che la trasformazione bilineare [9].

5.2.1 Filtri numerici Butterworth

I filtri di Butterworth sono definiti dalla proprietà che la risposta in ampiezza è massimamente piatta nella banda passante; per un filtro passa-basso dell' N -mo ordine, ciò significa che le prime $2N-1$ derivate del quadrato del modulo della risposta sono nulle in $\Omega = 0$. Un'altra proprietà è che l'approssimazione è monotona sia nella banda passante che nella banda oscura. La risposta (in modulo al quadrato) di un filtro analogico di Butterworth è della forma

$$|H_a(j\Omega)|^2 = \frac{1}{1 + (j\Omega/j\Omega_c)^{2N}} \quad (5.25)$$

schematizzata in fig. 5.10.

Al crescere del parametro N nell'espressione (5.25), la caratteristica del filtro diventa più ripida: essa rimane, cioè, prossima all'unità su un tratto sempre maggiore della banda passante e va sempre più rapidamente

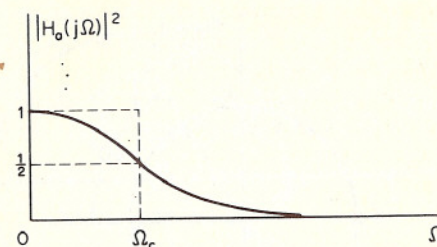


Fig. 5.10 Funzione modulo quadrato della risposta in frequenza di un filtro analogico di Butterworth.

a zero nella banda oscura, anche se, per la particolare forma della (5.25), il modulo continua a valere $1/\sqrt{2}$ alla frequenza di taglio Ω_c . La dipendenza della caratteristica del filtro Butterworth dal parametro N è indicata in fig. 5.11.

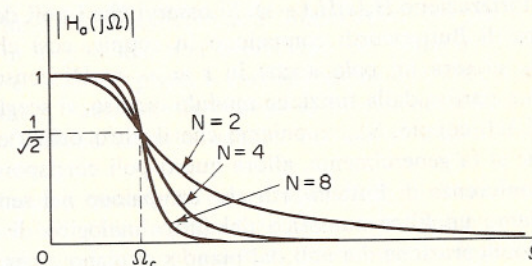


Fig. 5.11 Dipendenza del modulo della risposta di un filtro Butterworth dall'ordine N .

Dalla funzione modulo quadro (5.25) si può osservare che $H_a(s)H_a(-s)$ deve essere della forma

$$H_a(s)H_a(-s) = \frac{1}{1 + (s/j\Omega_c)^{2N}} \quad (5.26)$$

Le radici del polinomio a denominatore (i poli della funzione modulo quadro) cadono quindi in

$$s_p = (-1)^{1/2N} (j\Omega_c)$$

Perciò ci sono $2N$ poli equispaziati in angolo su una circonferenza di raggio Ω_c nel piano s . I poli sono disposti simmetricamente rispetto all'asse immaginario. Nessun polo può cadere sull'asse immaginario, mentre sull'asse reale ne cade uno per N dispari, ma non per N pari. La spaziatura angolare tra i poli lungo la circonferenza è di π/N radianti. Per esempio, per $N = 3$, i poli saranno spazati di $\pi/3$, cioè di 60 gradi, come indicato in fig. 5.12. Per determinare la funzione di trasferimento del filtro analogico da associare con la funzione modulo quadro di Butterworth, occorre

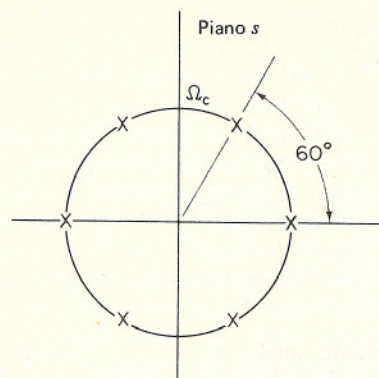


Fig. 5.12 Disposizione dei poli nel piano s per un filtro di Butterworth del terzo ordine.

eseguire la fattorizzazione $H_a(s)H_a(-s)$. Si osservi che i poli della funzione modulo quadro di Butterworth compaiono in coppie, così che se c'è un polo in $s = s_p$, ci sarà un polo anche in $s = -s_p$. Di conseguenza, per costruire $H_a(s)$ a partire dalla funzione modulo quadro, si sceglierà un polo da ciascuna di tali coppie. Se imponiamo che il filtro debba essere stabile e causale, come si fa generalmente, allora questi poli corrispondono a quei poli della circonferenza di Butterworth che compaiono nel semipiano sinistro. Se ricaviamo un filtro numerico dal filtro analogico di Butterworth mappando la configurazione dei poli dal piano s al piano z per mezzo della trasformazione bilineare, allora nel piano z la corrispondente funzione modulo quadro avrà $2N$ zeri in $z = -1$. La circonferenza di Butterworth nel piano s corrisponde ad una circonferenza nel piano z , in quanto la trasformazione bilineare è una trasformazione conforme. Tuttavia, la circonferenza di Butterworth nel piano z non è centrata nell'origine. Essa è rappresentata in fig. 5.13.

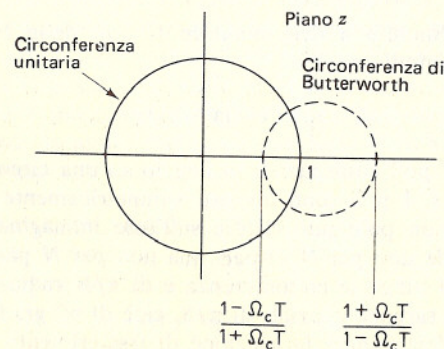


Fig. 5.13 Circonferenza di Butterworth trasformata con la trasformazione bilineare.

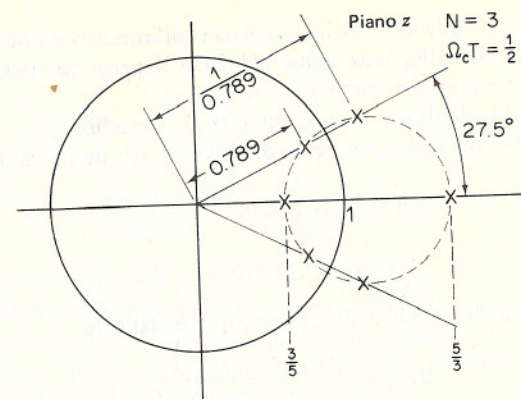


Fig. 5.14 Disposizione dei poli nel piano z per un filtro di Butterworth del terzo ordine trasformato con la trasformazione bilineare.

Mentre lungo la circonferenza di Butterworth del piano s i poli erano equispaziati in angolo, ciò non è più vero nel piano z : infatti, le coppie di poli in s_p e $-s_p$ nel piano s corrispondono alle coppie di poli in z_p e $1/z_p$ nel piano z . Facendo riferimento al precedente esempio con $N = 3$, i poli della funzione modulo quadro nel piano z sono rappresentati in fig. 5.14.

Generalmente, quando si progetta un filtro di Butterworth usando la trasformazione bilineare, la procedura più conveniente non è quella di localizzare direttamente i poli nel piano z , quanto quella di determinare innanzitutto la posizione dei poli nel piano s e poi mappare nel piano z , con la trasformazione bilineare, quei poli che si trovano nel semipiano sinistro.

Come esempio di progetto di un filtro numerico passa-basso di Butterworth, si assuma che sia richiesto un filtro con modulo della risposta in banda passante costante entro 1 dB per frequenze inferiori a 0.2π e con attenuazione in banda oscura maggiore di 15 dB per frequenze comprese tra 0.3π e π . Perciò, se si normalizza all'unità il modulo in banda passante per $\omega = 0$, allora le specifiche sono espresse da

$$20 \log_{10} |H(e^{j\omega})| \geq -1 \quad \text{e} \quad 20 \log_{10} |H(e^{j\omega})| \leq -15$$

A partire da queste specifiche, progetteremo ora un filtro numerico da un filtro Butterworth analogico utilizzando sia la tecnica dell'invarianza all'impulso, sia la trasformazione bilineare. Queste stesse specifiche saranno utilizzate per i successivi esempi degli altri metodi di approssimazione.

Progetto con l'invarianza all'impulso. Per progettare il filtro desiderato applicando l'invarianza all'impulso su un filtro di Butterworth analogico, bisogna per prima cosa trasformare le specifiche in termini della frequenza analogica. Si ricorda che, in assenza di *aliasing*, l'invarianza all'impulso corrisponde ad una trasformazione lineare dalla frequenza analogica alla frequenza numerica. Un'ipotesi di lavoro utile ai fini del progetto di

un filtro numerico sulla base dell'invarianza all'impulso è che l'effetto dell'*aliasing* sia trascurabile. Una volta concluso il progetto, potranno essere valutate le prestazioni del filtro risultante.

Nel progetto richiesto, assumeremo per comodità $T = 1$. Perciò la funzione $|H_a(j\Omega)|^2$ del filtro analogico di Butterworth desiderato deve essere tale che valga

$$20 \log_{10} |H_a(j.2\pi)| \geq -1$$

$$20 \log_{10} |H_a(j.3\pi)| \leq -15$$

Siccome la forma di un filtro di Butterworth è data da

$$|H_a(j\Omega)|^2 = \frac{1}{1 + (\Omega/\Omega_c)^{2N}}$$

il progetto del filtro consisterà essenzialmente nella determinazione dei parametri N e Ω_c che soddisfano le specifiche richieste. Determiniamo dapprima questi parametri in modo da soddisfare le specifiche con il segno di uguaglianza, così che

$$1 + \left(\frac{0.2\pi}{\Omega_c}\right)^{2N} = 10^{0.1} \quad (5.27)$$

e

$$1 + \left(\frac{0.3\pi}{\Omega_c}\right)^{2N} = 10^{1.5} \quad (5.28)$$

La soluzione di queste due equazioni fornisce i valori $N = 5.8858$ e $\Omega_c = 0.70474$. Il parametro N , però, deve essere intero e di conseguenza, per soddisfare le specifiche, va arrotondato al primo intero superiore, per cui $N = 6$. Ora però non è più possibile soddisfare con il segno di uguaglianza le specifiche sia sulla banda passante che su quella oscura: avendo arrotondato N per eccesso, si ha un miglioramento delle prestazioni del filtro e il modo in cui tale miglioramento si divide tra banda passante e banda oscura dipende dal valore di Ω_c . Sostituendo $N = 6$ nell'equazione (5.27), si ottiene $\Omega_c = 0.7032$. Con questo valore è soddisfatta esattamente la specifica sulla banda passante, e in eccesso quella sulla banda oscura (per il filtro analogico). Ciò comporta un ridotto effetto di *aliasing* sul filtro numerico. Con questo valore di Ω_c e con $N = 6$, ci sono tre coppie di poli nel semipiano sinistro del piano s , di coordinate:

$$\text{coppia 1: } -0.1820 \pm j 0.6792$$

$$\text{coppia 2: } -0.4972 \pm j 0.4972$$

$$\text{coppia 3: } -0.6792 \pm j 0.1820$$

per cui si può scrivere

$$H_a(s) = \frac{0.12093}{(s^2 + 0.3640s + 0.4945)(s^2 + 0.9945s + 0.4945)(s^2 + 1.3585s + 0.4945)}$$

Se si esprime $H_a(s)$ come espansione in fratti semplici e si applica la trasformazione (5.11), la funzione di trasferimento risultante per il filtro numerico sarà

$$H(z) = \frac{0.2871 - 0.4466z^{-1}}{1 - 0.1297z^{-1} + 0.6949z^{-2}} + \frac{-2.1428 + 1.1454z^{-1}}{1 - 1.0691z^{-1} + 0.3699z^{-2}} + \frac{1.8558 - 0.6304z^{-1}}{1 - 0.9972z^{-1} + 0.2570z^{-2}}$$

Come è evidente dalla sua espressione matematica, la funzione di trasferimento risultante dall'applicazione del metodo dell'invarianza all'impulso può essere realizzata direttamente in forma parallela. Se si preferisce la forma diretta o quella in cascata, i termini del secondo ordine separati devono essere combinati in maniera opportuna.

La risposta in frequenza del sistema trovato è riportata in fig. 5.15.

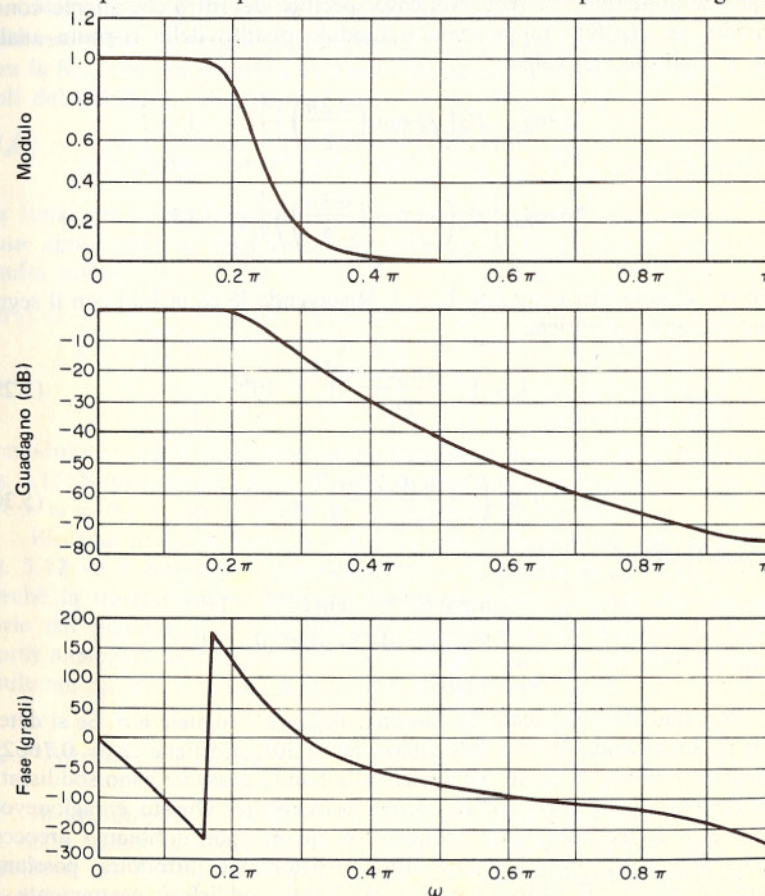


Fig. 5.15 Risposta in frequenza del filtro Butterworth del sesto ordine trasformato secondo l'invarianza all'impulso.

Si ricorda che il filtro è stato progettato in modo da soddisfare esattamente la specifica sulla banda passante e in eccesso quella sulla banda oscura, e ciò si è effettivamente ottenuto. Questa è un'indicazione che il filtro analogico era sufficientemente limitato in banda per non comportare problemi di *aliasing*. Altre volte le cose non vanno così; se il filtro numerico risultante non soddisfa le specifiche, si può riprovare con un filtro di ordine superiore oppure, mantenendo l'ordine fissato, con un diverso ritocco dei parametri del filtro.

Progetto per mezzo della trasformazione bilineare. Come si è visto precedentemente, nel progetto con la trasformazione bilineare le specifiche sulle frequenze numeriche devono essere trasferite in campo analogico, in modo che, con la distorsione di frequenza inerente alla trasformazione bilineare, le frequenze analogiche critiche si mappino correttamente nelle frequenze numeriche critiche. Nel caso specifico del filtro che stiamo considerando, se $|H_a(j\Omega)|^2$ rappresenta il modulo quadro della risposta analogica, si richiede che valga

$$20 \log_{10} \left| H_a \left(j2 \tan \left(\frac{0.2\pi}{2} \right) \right) \right| \geq -1$$

e

$$20 \log_{10} \left| H_a \left(j2 \tan \left(\frac{0.3\pi}{2} \right) \right) \right| \leq -15$$

dove si è assunto per comodità $T = 1$. Risolvendo le equazioni con il segno di uguaglianza, si ottiene

$$1 + \left(\frac{2 \tan(0.1\pi)}{\Omega_c} \right)^{2N} = 10^{0.1} \quad (5.29)$$

e

$$1 + \left(\frac{2 \tan(0.15\pi)}{\Omega_c} \right)^{2N} = 10^{1.5} \quad (5.30)$$

e perciò si ricava

$$N = \frac{1 \log [(10^{1.5} - 1)/(10^{0.1} - 1)]}{2 \log [\tan(0.15\pi)/\tan(0.1\pi)]} = 5.30466$$

Per soddisfare le specifiche bisogna scegliere N uguale a 6. Se si determina Ω_c sostituendo $N = 6$ nell'equazione (5.30), si ottiene $\Omega_c = 0.76622$. Con questo valore di Ω_c , le specifiche sulla banda passante sono soddisfatte in eccesso e quelle sulla banda oscura esattamente. Questo è ragionevole nel caso della trasformazione bilineare, in quanto non dobbiamo preoccuparci dell'*aliasing*. Infatti, grazie alla pre-distorsione introdotta, possiamo essere sicuri che il filtro numerico risultante soddisferà esattamente la specifica al limite della banda oscura.

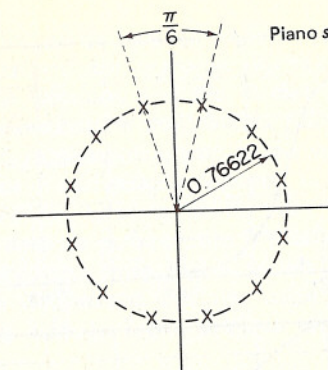


Fig. 5.16 Disposizione dei poli nel piano s per un filtro di Butterworth del sesto ordine.

Nel piano s i 12 poli della funzione modulo quadro sono distribuiti uniformemente in angolo sulla circonferenza di raggio 0.76622, come mostra la fig. 5.16. La funzione di trasferimento nel piano s corrispondente ai poli del semipiano sinistro è

$$H_a(s) = \frac{0.20238}{(s^2 + 0.396s + 0.5871)(s^2 + 1.083s + 0.5871)(s^2 + 1.4802s + 0.5871)}$$

La funzione di trasferimento $H(z)$ del filtro numerico viene quindi ottenuta applicando la trasformazione bilineare a $H_a(s)$ con T unitario, e risulta quindi

$$H(z) = \frac{0.0007378(1 + z^{-1})^6}{(1 - 1.2686z^{-1} + 0.7051z^{-2})(1 - 1.0106z^{-1} + 0.3583z^{-2})} \times \frac{1}{(1 - 0.9044z^{-1} + 0.2155z^{-2})}$$

Il modulo e la fase della risposta in frequenza numerica sono riportati in fig. 5.17. Si può notare che a $\omega = 0.2\pi$ il modulo è ridotto di 0.5632 dB e che a $\omega = 0.3\pi$ esso è diminuito esattamente di 15 dB.

Bisogna anche mettere in evidenza che il diagramma del modulo in fig. 5.17 va a zero molto più rapidamente di quello di fig. 5.15. Questo perché la trasformazione bilineare fa corrispondere l'intero asse immaginario del piano s al circolo unitario. Perciò, siccome il filtro di Butterworth analogico ha uno zero del sesto ordine in $s = \infty$, il filtro numerico risultante ha uno zero del sesto ordine in $z = -1$.

5.2.2 Filtri numerici Chebyshev

In un filtro di Butterworth la caratteristica di frequenza è monotona sia nella banda passante che nella banda oscura. Di conseguenza, se le specifiche del filtro sono in termini, ad esempio, di massimo errore di approssimazione nella banda passante, tali specifiche vengono soddisfatte

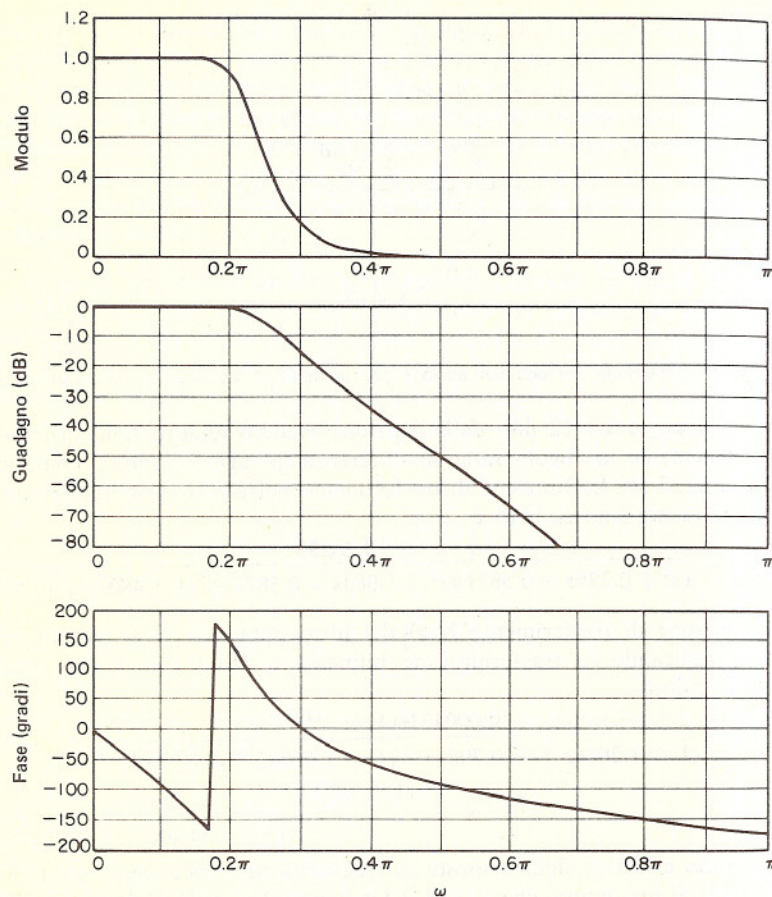


Fig. 5.17 Risposta in frequenza del filtro Butterworth del sesto ordine trasformato con la trasformazione bilineare.

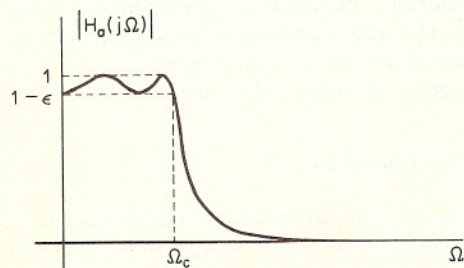


Fig. 5.18 Approssimazione di Chebyshev per un filtro passa-basso.

con una precisione che eccede tanto più quella richiesta quanto più ci si avvicina alla frequenza zero. Un approccio più efficiente, che di solito conduce a filtri di ordine inferiore, è quello di distribuire la precisione dell'approssimazione uniformemente nella banda passante o in quella oscura o in entrambe. Si può raggiungere questo scopo scegliendo un'approssimazione con caratteristica di oscillazione uniforme (« equiripple ») invece che con caratteristica monotona. La classe dei filtri di Chebyshev ha la proprietà che il modulo della risposta in frequenza è a oscillazione uniforme nella banda passante e monotona nella banda oscura oppure monotona nella banda passante ed a oscillazione uniforme nella banda oscura. Il primo caso è illustrato in fig. 5.18. La forma analitica per il quadrato del modulo della risposta è

$$|H_a(\Omega)|^2 = \frac{1}{1 + \varepsilon^2 V_N^2(\Omega/\Omega_c)} \quad (5.31)$$

dove $V_N(x)$ è il polinomio di Chebyshev di ordine N definito come

$$V_N(x) = \cos(N \cos^{-1} x) \quad (5.32)$$

Ad esempio, risulta, per $N = 0$, $V_N(x) = 1$; per $N = 1$, $V_N(x) = \cos(\cos^{-1} x) = x$; per $N = 2$, $V_N(x) = \cos(2\cos^{-1} x) = 2x^2 - 1$, etc.

Dalla definizione dei polinomi di Chebyshev (5.32) è immediato ricavare una formula ricorsiva per ottenere $V_{N+1}(x)$ da $V_N(x)$ e $V_{N-1}(x)$, applicando delle identità trigonometriche alla (5.32): il risultato è

$$V_{N+1}(x) = 2xV_N(x) - V_{N-1}(x) \quad (5.33)$$

Dalla definizione (5.32) notiamo che $V_N^2(x)$ varia tra zero e uno per x tra zero e uno. Per x maggiore di uno, $\cos^{-1} x$ è immaginario, per cui $V_N(x)$ si comporta come un coseno iperbolico e di conseguenza è monotono crescente. Tornando allora all'espressione (5.31), $|H_a(\Omega)|^2$ oscilla tra 1 e $1/(1 + \varepsilon^2)$ per $0 \leq \Omega/\Omega_c \leq 1$ e decresce monotonicamente per $\Omega/\Omega_c > 1$. Per specificare il filtro sono necessari tre parametri: ε , Ω_c ed N . In una tipica situazione di progetto, ε è determinato dall'oscillazione ("ripple") ammessa nella banda passante e Ω_c dalla frequenza di taglio desiderata. L'ordine N viene poi scelto in modo da soddisfare le specifiche relative alla banda oscura.

I poli del filtro di Chebyshev sono disposti su un'ellisse nel piano s [2, 7, 8]. Con riferimento alla fig. 5.19, l'ellisse è definita da due circonferenze corrispondenti all'asse maggiore e all'asse minore dell'ellisse medesima. Il raggio dell'asse minore è $a\Omega_c$, dove

$$a = \frac{1}{2}(\alpha^{1/N} - \alpha^{-1/N}) \quad (5.34)$$

con

$$\alpha = \varepsilon^{-1} + \sqrt{1 + \varepsilon^{-2}} \quad (5.35)$$

Il raggio dell'asse maggiore è $b\Omega_c$, dove

$$b = \frac{1}{2}(\alpha^{1/N} + \alpha^{-1/N}) \quad (5.36)$$

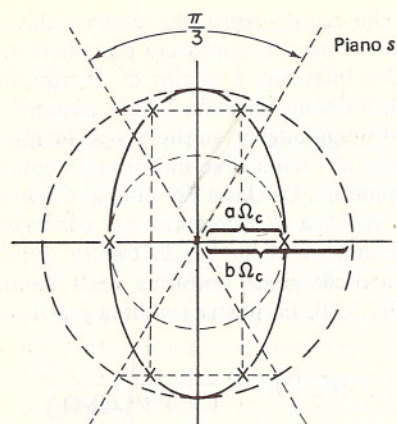


Fig. 5.19 Posizione dei poli per un filtro di Chebyshev del terzo ordine.

Per collocare i poli del filtro di Chebyshev sull'ellisse, dapprima identifichiamo sulle circonferenze maggiore e minore i punti con spaziatura angolare uniforme pari a π/N , in modo che i punti siano in posizioni simmetriche rispetto all'asse immaginario, che nessun punto cada sull'asse immaginario e che un punto cada sull'asse reale per N dispari ma non per N pari. Questa suddivisione delle circonferenze maggiore e minore corrisponde esattamente al modo in cui si divide la circonferenza per individuare i poli di un filtro di Butterworth. I poli di un filtro di Chebyshev si trovano sull'ellisse, con ordinate uguali a quelle dei punti identificati sulla circonferenza maggiore e ascisse uguali a quelle dei punti identificati sulla circonferenza minore. In fig. 5.19 sono mostrati i poli per $N = 3$.

Come esempio di progetto di un filtro di Chebyshev, consideriamo le stesse specifiche usate per il filtro di Butterworth e confrontiamo di nuovo il progetto basato sull'invarianza all'impulso con il progetto che usa la trasformazione bilineare.

Progetto con l'invarianza all'impulso. Stiamo cercando un filtro di Chebyshev analogico la cui risposta in frequenza ha modulo quadro che soddisfa le specifiche

$$20 \log_{10} |H_a(j.2\pi)| \geq -1$$

e

$$20 \log_{10} |H_a(j.3\pi)| \leq -15$$

Sceghieremo i parametri di progetto in modo da soddisfare esattamente la specifica a 0.2π con una risposta in frequenza ad oscillazione uniforme tra $\Omega = 0$ e $\Omega = 0.2\pi$. Di conseguenza, sarà $\Omega_c = 0.2\pi$ e $\delta_1 = 10^{-0.05} \Rightarrow \epsilon = 0.50885$. Risulta, per $N = 3$, $20 \log_{10} |H_a(j.3\pi)| = -13.4189$, e per $N = 4$, $20 \log_{10} |H_a(j.3\pi)| = -21.5834$. Perciò scegliamo il valore maggiore di N . Per questo valore di N i parametri α , a e b sono $\alpha = 4.1702$,

$a = 0.3646$ e $b = 1.0644$. Quindi si ha

$$H_a(s) = \frac{0.038286}{(s^2 + 0.4233s + 0.1103)(s^2 + 0.1753s + 0.3894)}$$

La funzione di trasferimento del filtro numerico ottenuto usando l'invarianza all'impulso è

$$H(z) = \frac{0.08327 + 0.0239z^{-1}}{1 - 1.5658z^{-1} + 0.6549z^{-2}} - \frac{0.08327 + 0.0246z^{-1}}{1 - 1.4934z^{-1} + 0.8392z^{-2}}$$

Vale la pena di notare che per il fenomeno dell'*aliasing* l'attenuazione al limite della banda oscura, cioè per $\Omega = 0.3\pi$, è leggermente peggiore che non per il filtro analogico. Tuttavia, poiché il progetto analogico forniva un'attenuazione maggiore di quella richiesta, dovendo N essere scelto intero, il filtro numerico che ne risulta soddisfa le specifiche. Un grafico del modulo e della fase della risposta in frequenza risultante è mostrato in fig. 5.20.

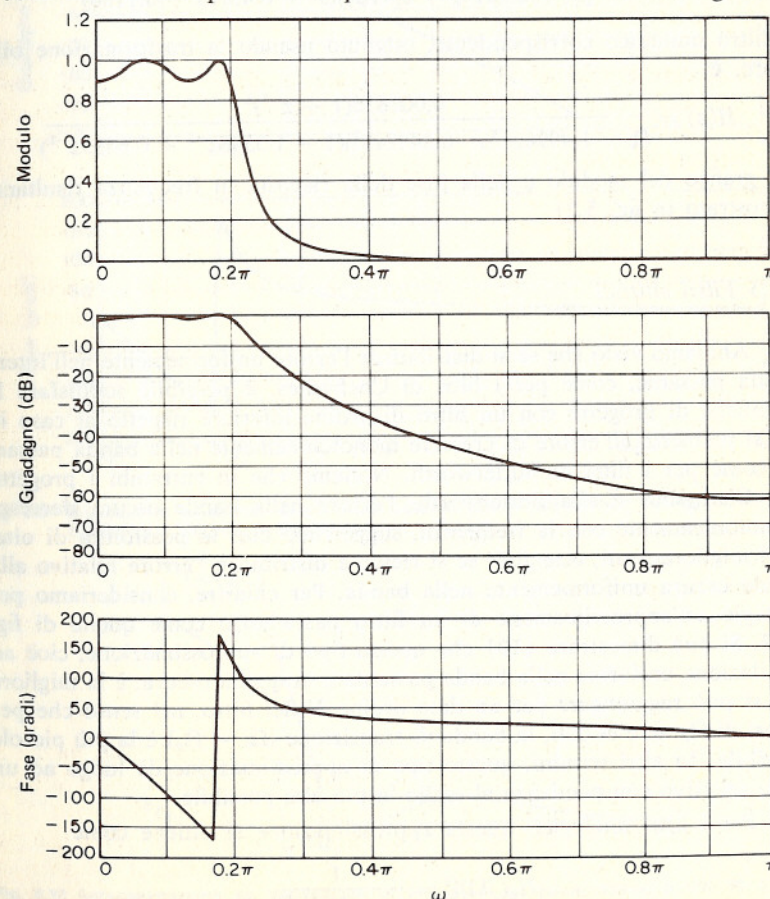


Fig. 5.20 Risposta in frequenza di un filtro passa-basso di Chebyshev del quarto ordine trasformato usando l'invarianza all'impulso.

Progetto per mezzo della trasformazione bilineare. In questo caso le specifiche relative al filtro analogico sono

$$20 \log_{10} \left| H_a \left(j2 \tan \left(\frac{0.2\pi}{2} \right) \right) \right| \geq -1$$

e

$$20 \log_{10} \left| H_a \left(j2 \tan \left(\frac{0.3\pi}{2} \right) \right) \right| \leq -15$$

Quindi il parametro Ω_c vale $\Omega_c = 2 \tan(0.2\pi/2)$ e, come nel caso precedente, $\delta_1 = 10^{-0.05} \Rightarrow \epsilon = 0.50885$. Il minimo valore intero di N che permette di soddisfare le specifiche relative alla banda oscura è $N = 4$. La funzione di trasferimento del filtro analogico che se ne ricava è

$$H_a(s) = \frac{0.04381}{(s^2 + 0.1814s + 0.4166)(s^2 + 0.4378s + 0.1180)}$$

Il filtro numerico corrispondente, ottenuto usando la trasformazione bilineare, è

$$H(z) = \frac{0.001836(1 + z^{-1})^4}{(1 - 1.4996z^{-1} + 0.8482z^{-2})(1 - 1.5548z^{-1} + 0.6493z^{-2})}$$

Un grafico del modulo e della fase della risposta in frequenza risultante è mostrato in fig. 5.21.

5.2.3 Filtri ellittici

Abbiamo visto che se si distribuisce l'errore uniformemente nell'intera banda passante, come per i filtri di Chebyshev, è possibile soddisfare le specifiche di progetto con un filtro di ordine inferiore rispetto al caso in cui si permette all'errore di crescere monotonicamente nella banda passante, come per i filtri di Butterworth. Notiamo che in entrambi i progetti, alla Chebyshev e alla Butterworth, l'errore nella banda oscura decresce monotonicamente con la frequenza, suggerendo così la possibilità di ulteriori miglioramenti, ottenibili se si riesce a distribuire l'errore relativo alla banda oscura uniformemente nella banda. Per chiarire, consideriamo per esempio un'approssimazione di un filtro passa-basso come quella di fig. 5.22. Si può dimostrare [10] che questo tipo di approssimazione, cioè ad oscillazione uniforme nella banda passante e in quella oscura, è la migliore che si può raggiungere per un dato ordine N del filtro, nel senso che per valori di Ω_p , δ_1 e δ_2 dati, la banda di transizione ($\Omega_s - \Omega_p$) è la più piccola possibile. In altri termini, questo tipo di approssimazione dà luogo ad un filtro selettivo con pendenza al taglio la più alta possibile.

Per i filtri analogici, questa approssimazione si ottiene come

$$|H_a(j\Omega)|^2 = \frac{1}{1 + \epsilon^2 U_N^2(\Omega)}$$

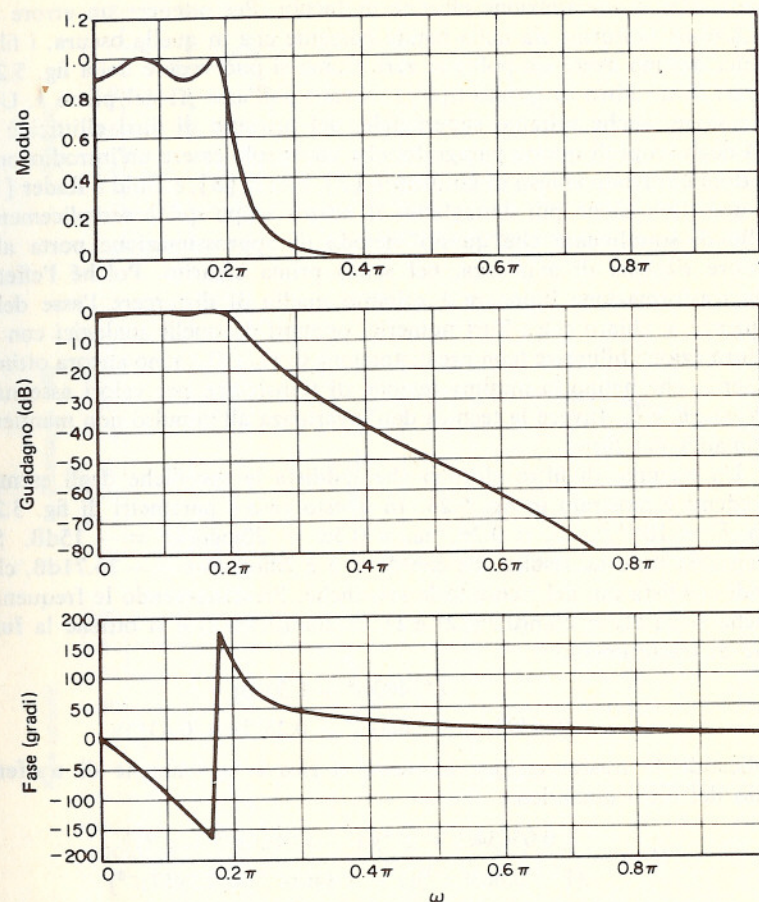


Fig. 5.21 Risposta in frequenza di un filtro passa-basso di Chebyshev del quarto ordine trasformato usando la trasformazione bilineare.

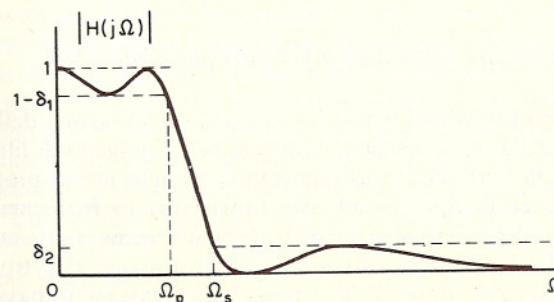


Fig. 5.22 Approssimazione ad oscillazione uniforme sia in banda passante che in banda oscura.

dove $U_N(\Omega)$ è una funzione ellittica di Jacobi. Per ottenere un errore ad oscillazione uniforme sia nella banda passante che in quella oscura, i filtri ellittici devono avere sia poli che zeri. Come si può vedere dalla fig. 5.22, gli zeri di un filtro di questo tipo si trovano sull'asse $j\Omega$ del piano s . Una discussione, anche soltanto superficiale, del progetto di filtri ellittici è al di là degli scopi di questo paragrafo, che vuole solo essere un'introduzione. Il lettore è rinviato ai testi di Guillemin [7], Storer [8], e Gold e Rader [1] per una discussione più dettagliata. Il nostro scopo qui è semplicemente quello di sottolineare che questo metodo di approssimazione porta alla migliore risposta di ampiezza, nel senso prima chiarito. Poiché l'effetto della trasformazione bilineare è soltanto quello di distorcere l'asse delle frequenze, è chiaro che i filtri numerici ottenuti da quelli analogici con la trasformazione bilineare (con pre-distorsione di Ω_p e Ω_s) sono ancora ottimi, nel senso che hanno la minima regione di transizione per valori assegnati di N , ω_p , δ_1 e δ_2 . Invece la tecnica dell'invarianza all'impulso non mantiene l'ottimalità del filtro.

Un esempio di filtro ellittico che soddisfa le specifiche degli esempi precedenti è mostrato in fig. 5.23. In questo caso i parametri di fig. 5.22 sono $\delta_1 = 10^{-0.05}$, $\omega_p = 0.2\pi$, $\omega_s = 0.3\pi$ e $20\log_{10}(\delta_2) = -15$ dB. Se fissiamo δ_1 , ω_p e ω_s , risulta che con $N = 3$ è $20\log_{10}(\delta_2) = -26.71$ dB, che quindi soddisfa più del richiesto le specifiche. Pre-distorcendo le frequenze critiche si ha $\Omega_p = 2\tan(0.2\pi/2)$ e $\Omega_s = 2\tan(0.3\pi/2)$ e si ottiene la funzione di trasferimento

$$H_a(s) = \frac{0.12460(s^2 + 1.3040)}{(0.6498s + 0.2448)(s^2 + 0.2521s + 0.4313)}$$

Applicando la trasformazione bilineare si ricava la funzione di trasferimento del filtro numerico

$$H(z) = \frac{0.05634(1 + z^{-1})(1 - 1.0166z^{-1} + z^{-2})}{(1 - 0.6830z^{-1})(1 - 1.4461z^{-1} + 0.7957z^{-2})}$$

Vediamo quindi che per l'esempio considerato il filtro ellittico è quello di ordine minimo che soddisfa le specifiche.

5.2.4 Trasformazioni di frequenza per filtri passa-basso IIR

Gli esempi precedenti hanno illustrato l'uso dei metodi dell'invarianza all'impulso e della trasformazione bilineare per il progetto di filtri numerici IIR a partire da funzioni di trasferimento analogiche aventi proprietà selettive in frequenza di tipo passa-basso. Le risposte in frequenza ideali dei quattro tipi di filtri selettivi in frequenza comunemente usati sono mostrate in fig. 5.24. Le fig. 5.24(a), (b), (c) e (d) presentano, rispettivamente, la risposta in frequenza ideale di filtri passa-basso, passa-alto, passa-banda ed elimina-banda. L'approccio tradizionale al progetto di questi filtri analogici selettivi in frequenza è quello di progettare dapprima un prototipo di filtro

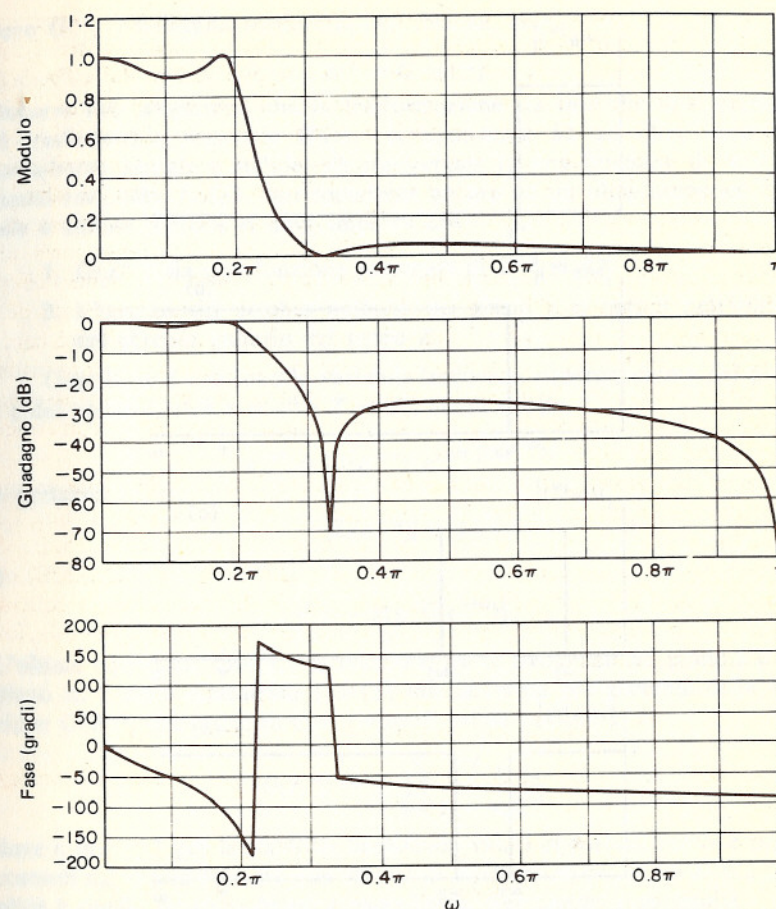


Fig. 5.23 Risposta in frequenza di un filtro ellittico del terzo ordine trasformato usando la trasformazione bilineare.

passa-basso con frequenza normalizzata e poi ricavare, con trasformazioni algebriche, dal prototipo passa-basso il filtro con la caratteristica desiderata, passa-basso, passa-alto, passa-banda o elimina-banda [7]. Nel caso dei filtri numerici selettivi in frequenza, possiamo progettare un filtro selettivo analogico del tipo desiderato e poi trasformarlo in filtro numerico. Uno svantaggio di questo procedimento è che non possiamo trasformare filtri passa-alto o elimina-banda usando la tecnica dell'invarianza all'impulso, a causa della distorsione che introdurrebbe l'aliasing. Un metodo alternativo è quello di progettare un prototipo di filtro numerico passa-basso e poi eseguire su di questo una trasformazione algebrica per ottenere il filtro selettivo in frequenza richiesto [11-13]. Questo procedimento può essere applicato indipendentemente dal metodo di progetto seguito per ricavare il filtro numerico passa-basso.

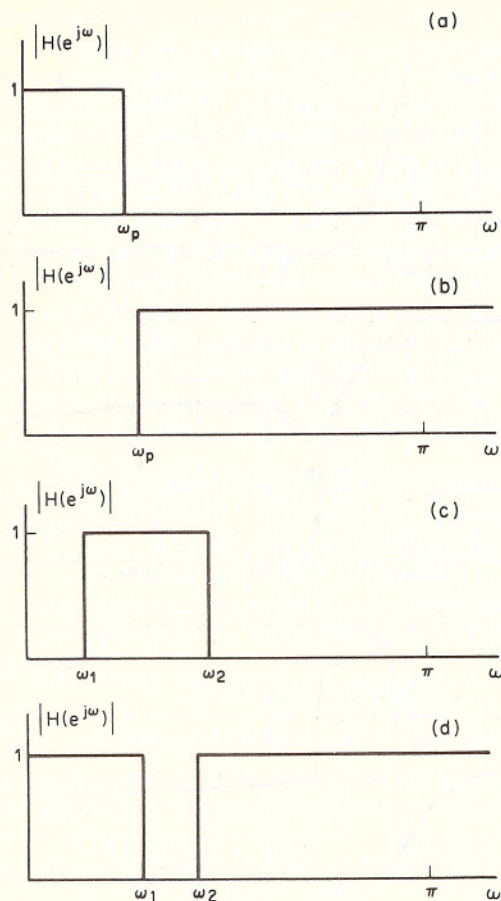


Fig. 5.24 Risposte in frequenza di filtri ideali (a) passa-basso, (b) passa-alto, (c) passa-banda, (d) elimina-banda.

I filtri selettivi in frequenza del tipo passa-basso, passa-alto, passa-banda ed elimina-banda possono essere ottenuti da un filtro passa-basso numerico mediante trasformazioni razionali molto simili alla trasformazione bilineare che abbiamo usato per passare da funzioni di trasferimento analogiche a funzioni di trasferimento numeriche. Per chiarire come questo può essere fatto, associamo la variabile complessa z alla funzione di trasferimento passa-basso, $H_l(z)$, e la variabile complessa Z alla funzione di trasferimento desiderata $H_d(Z)$. Definiamo poi una trasformazione dal piano z al piano Z della forma

$$z^{-1} = G(Z^{-1})$$

in modo che risulti

$$H_d(Z) = H_l(G^{-1}(Z^{-1}))$$

dove $G^{-1}(\)$ indica la trasformazione inversa, cioè

$$Z^{-1} = G^{-1}(z^{-1})$$

Occorre fare attenzione che la trasformazione sia tale che una funzione di trasferimento razionale $H_l(z)$, corrispondente ad un filtro numerico passa-basso causale e stabile, sia trasformata in una funzione di trasferimento razionale, $H_d(Z)$, corrispondente ancora ad un filtro numerico causale e stabile. Perciò, si deve imporre che

1. $G(Z^{-1})$ sia una funzione razionale di Z^{-1} (o Z).
2. L'interno del cerchio unitario del piano z si mappi nell'interno del cerchio unitario del piano Z .

Quindi, se θ e ω sono le variabili frequenza rispettivamente del piano z e del piano Z , cioè $z = e^{j\theta}$ e $Z = e^{j\omega}$, allora risulta

$$e^{-j\theta} = |G(e^{-j\omega})| e^{j \arg [G(e^{-j\omega})]}$$

e pertanto

$$|G(e^{-j\omega})| = 1$$

e

$$\theta = -\arg [G(e^{-j\omega})]$$

L'ultima equazione specifica la relazione tra le frequenze del piano z e del piano Z . È stato dimostrato [11-13] che la forma più generale della funzione $G(Z^{-1})$ che soddisfa tutti i vincoli prima imposti è

$$G(Z^{-1}) = \pm \prod_{k=1}^N \frac{Z^{-1} - \alpha_k}{1 - \alpha_k Z^{-1}}$$

dove è $|\alpha_k| < 1$ per la stabilità. Scegliendo valori opportuni per N e per le costanti α_k , si può ottenere una varietà di trasformazioni, di cui la più semplice è quella da passa-basso a passa-basso. Per questo caso risulta

$$z^{-1} = G(Z^{-1}) = \frac{Z^{-1} - \alpha}{1 - \alpha Z^{-1}}$$

Sostituendo $z = e^{j\theta}$ e $Z = e^{j\omega}$, otteniamo

$$e^{-j\theta} = \frac{e^{-j\omega} - \alpha}{1 - \alpha e^{-j\omega}}$$

da cui si può dimostrare che è

$$\omega = \arctan \left[\frac{(1 - \alpha^2) \sin \theta}{2\alpha + (1 + \alpha^2) \cos \theta} \right]$$

Questa relazione è illustrata graficamente in fig. 5.25 per diversi valori di α . Anche se risulta evidente dalla fig. 5.25 una distorsione della scala delle frequenze (ad eccezione del caso $\alpha = 0$), se il sistema originale ha una risposta in frequenza passa-basso costante a tratti con frequenza di taglio

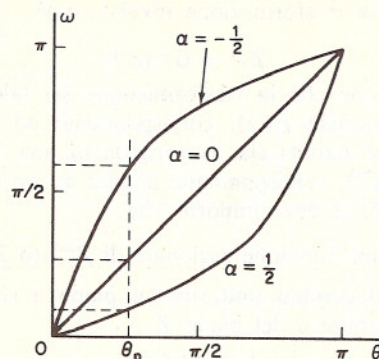


Fig. 5.25 Distorsione della scala delle frequenze nella trasformazione da passa-basso a passa-basso.

θ_p , allora il sistema trasformato avrà una risposta passa-basso simile, con frequenza di taglio ω_p determinata dalla scelta di α . Esplicitando α in termini di θ_p e ω_p , si ottiene

$$\alpha = \frac{\sin((\theta_p - \omega_p)/2)}{\sin((\theta_p + \omega_p)/2)}$$

Quindi, per usare questi risultati allo scopo di ricavare un filtro passa-basso $H_d(Z)$ con frequenza di taglio ω_p a partire da un passa-basso $H_l(z)$ già disponibile, con frequenza di taglio θ_p , occorre usare la relazione precedente per determinare α nell'espressione

$$H_d(Z) = H_l(z) \Big|_{z^{-1} = (Z^{-1} - \alpha)/(1 - \alpha Z^{-1})}$$

In maniera analoga possono essere ricavate le trasformazioni da passa-basso a passa-alto, a passa-banda ed elimina-banda. Queste trasformazioni sono riassunte in tab. 5.1 [11-13].

Come esempio d'uso di queste trasformazioni, ricaviamo un filtro passa-alto dal filtro passa-basso di Chebyshev del par. 5.2.2 Ricordiamo che la frequenza di taglio del filtro passa-basso era $\theta_p = 0.2\pi$. Usando la trasformazione bilineare avevamo ottenuto

$$H_l(z) = \frac{0.001836(1 + z^{-1})^4}{(1 - 1.5548z^{-1} + 0.6493z^{-2})(1 - 1.4996z^{-1} + 0.8482z^{-2})}$$

La risposta in frequenza di questo filtro è mostrata in fig. 5.21. Supponiamo di volere un filtro passa-alto con frequenza di taglio $\omega_p = 0.6\pi$. Dalla tab. 5.1 si ricava

$$\alpha = -\frac{\cos[(0.6\pi + 0.2\pi)/2]}{\cos[(0.6\pi - 0.2\pi)/2]} = -0.38197$$

Tab. 5.1 Trasformazioni da un prototipo di filtro numerico passa-basso con frequenza di taglio θ_p .

Tipo di filtro	Trasformazione	Formule di progetto
Passa-basso	$z^{-1} = \frac{Z^{-1} - \alpha}{1 - \alpha Z^{-1}}$	$\alpha = \frac{\sin\left(\frac{\theta_p - \omega_p}{2}\right)}{\sin\left(\frac{\theta_p + \omega_p}{2}\right)}$ $\omega_p = \text{freq. di taglio desiderata}$
Passa-alto	$z^{-1} = \frac{Z^{-1} + \alpha}{1 + \alpha Z^{-1}}$	$\alpha = -\frac{\cos\left(\frac{\omega_p + \theta_p}{2}\right)}{\cos\left(\frac{\omega_p - \theta_p}{2}\right)}$ $\omega_p = \text{freq. di taglio desiderata}$
Passa-banda	$z^{-1} = \frac{Z^{-2} - \frac{2\alpha k}{k+1} Z^{-1} + \frac{k-1}{k+1}}{\frac{k-1}{k+1} Z^{-2} - \frac{2\alpha k}{k+1} Z^{-1} + 1}$	$\alpha = \frac{\cos\left(\frac{\omega_2 + \omega_1}{2}\right)}{\cos\left(\frac{\omega_2 - \omega_1}{2}\right)}$ $k = \cot\left(\frac{\omega_2 - \omega_1}{2}\right) \tan \frac{\theta_p}{2}$ $\omega_2, \omega_1 = \text{freq. di taglio superiore e inferiore desiderate}$
Elimina-banda	$z^{-1} = \frac{Z^{-2} - \frac{2\alpha}{1+k} Z^{-1} + \frac{1-k}{1+k}}{\frac{1-k}{1+k} Z^{-2} - \frac{2\alpha}{1+k} Z^{-1} + 1}$	$\alpha = \frac{\cos\left(\frac{\omega_2 + \omega_1}{2}\right)}{\cos\left(\frac{\omega_2 - \omega_1}{2}\right)}$ $k = \tan\left(\frac{\omega_2 - \omega_1}{2}\right) \tan \frac{\theta_p}{2}$ $\omega_2, \omega_1 = \text{freq. di taglio superiore e inferiore desiderate}$

Perciò, usando la trasformazione da passa-basso a passa-alto indicata in tab. 5.1, otteniamo

$$H_d(Z) = H_l(z) \Big|_{z^{-1} = (Z^{-1} - 0.38197)/(1 - 0.38197Z^{-1})} = \frac{0.02426(1 - Z^{-1})^4}{(1 - 1.0416Z^{-1} + 0.4019Z^{-2})(1 - 0.5561Z^{-1} + 0.7647Z^{-2})}$$

La risposta in frequenza di questo sistema è rappresentata in fig. 5.26. Si noti che, a parte una certa distorsione della scala delle frequenze, la risposta passa-alto assomiglia molto alla risposta passa-basso traslata di π in frequenza. Notiamo anche che lo zero del quarto ordine in $z = -1$ del filtro passa-basso compare adesso in $Z = 1$ per il filtro passa-alto.

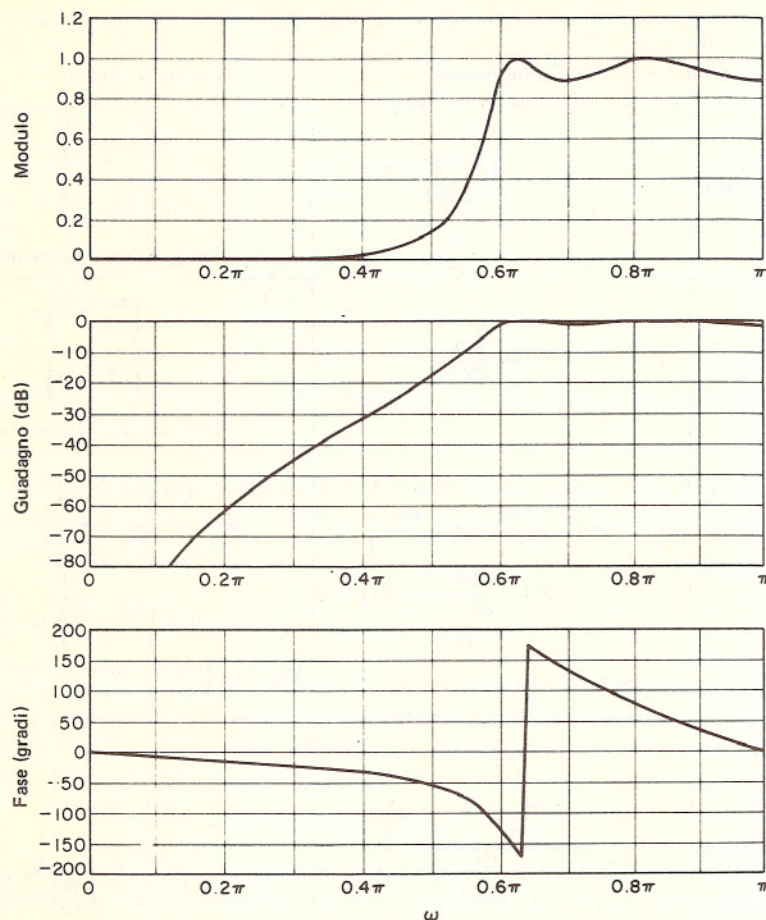


Fig. 5.26 Risposta in frequenza di un filtro passa-alto ottenuto con una trasformazione di frequenza.

5.3 PROGETTO DI FILTRI NUMERICI IIR ASSISTITO DA CALCOLATORE

Nei paragrafi precedenti abbiamo visto che si possono progettare filtri numerici trasformando un filtro analogico opportunamente progettato. Questo approccio è conveniente quando si può trarre profitto da progetti analogici che sono dati in termini di formule o di tabelle di progetto complete: ad esempio, i filtri selettivi in frequenza di tipo Butterworth, Chebyshev o ellittici. In generale, però, non esistono procedimenti analitici per il progetto di filtri analogici o numerici che soddisfino specifiche arbitrarie sulla risposta in frequenza o altri tipi di specifiche. Per questi casi più generali, sono state sviluppate delle tecniche di progetto di natura algoritmica e che

si basano generalmente sull'uso di un calcolatore per risolvere sistemi di equazioni lineari o non lineari. In molti casi le tecniche di progetto assistite da calcolatore si applicano egualmente bene al progetto di filtri sia numerici che analogici, effettuando soltanto qualche piccola modifica. Di conseguenza non si ottiene alcun vantaggio nell'effettuare prima un progetto analogico per trasformarlo poi in un progetto numerico.

Esistono numerose tecniche di progettazione assistita da calcolatore che consentono di approssimare una risposta in frequenza arbitraria. In questo paragrafo discuteremo alcune di queste procedure così da illustrare le possibilità del progetto assistito da calcolatore di filtri numerici IIR. Concentreremo l'attenzione sul modo in cui si impostano le equazioni di progetto anziché sui dettagli delle procedure numeriche richieste per ottenere la soluzione.

5.3.1 Minimizzazione dell'errore quadratico medio

Steiglitz [14,15] ha proposto una tecnica di progetto per filtri IIR basata sulla minimizzazione dell'errore quadratico nel dominio della frequenza. Essa richiede che la risposta in frequenza desiderata $H_d(e^{j\omega})$ sia assegnata per un insieme discreto di frequenze $\{\omega_i\}$, $i = 1, \dots, M$. L'errore quadratico medio a queste frequenze è definito come

$$E = \sum_{i=1}^M [|H(e^{j\omega_i})| - |H_d(e^{j\omega_i})|]^2 \quad (5.37)$$

Si assume che la funzione di trasferimento del filtro sia della forma

$$H(z) = A \prod_{k=1}^K \frac{1 + a_k z^{-1} + b_k z^{-2}}{1 + c_k z^{-1} + d_k z^{-2}} = AG(z) \quad (5.38)$$

Viene scelta la forma in cascata a causa della sua relativamente bassa sensibilità alle variazioni dei coefficienti ed anche perché conveniente nel calcolo delle derivate richieste dalla procedura di ottimizzazione.

L'errore, espresso dalla (5.37), può essere pensato come una funzione dei parametri $(a_1, b_1, c_1, d_1, a_2, b_2, \dots, d_K, A)$. Poiché desideriamo trovare i valori di questi parametri che minimizzano l'errore E , effettuiamo l'operazione di derivazione parziale rispetto a ciascuno dei parametri ed eguagliamo tali derivate a zero in modo da ottenere $4K+1$ equazioni in $4K+1$ incognite.

L'equazione per A è particolarmente semplice poiché

$$\frac{\partial E}{\partial |A|} = \sum_{i=1}^M \{2[|A| \cdot |G(e^{j\omega_i})| - |H_d(e^{j\omega_i})|] |G(e^{j\omega_i})|\} = 0$$

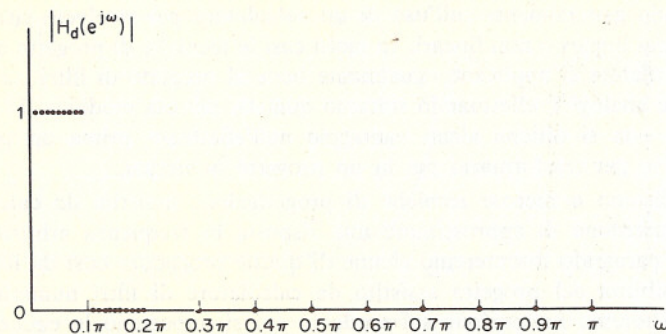


Fig. 5.27 Valori preassegnati della risposta in frequenza per un esempio sull'uso della procedura di progetto di Steiglitz.

Risolvendo questa equazione rispetto ad A si ha³

$$|A| = \frac{\sum_{i=1}^M |G(e^{j\omega_i})| \cdot |H_d(e^{j\omega_i})|}{\sum_{i=1}^M |G(e^{j\omega_i})|^2} \quad (5.39)$$

Differenziando rispetto ai rimanenti $4K$ parametri incogniti rappresentati dal vettore

$$\Phi = [a_1, b_1, c_1, d_1, a_2, b_2, \dots, d_K]$$

si ottengono $4K$ equazioni non lineari

$$\frac{\partial E(\Phi, A)}{\partial \phi_n} = 0, \quad n = 1, 2, \dots, 4K$$

dove ϕ_n rappresenta la n -ma componente di Φ . Queste equazioni possono essere risolte algebricamente usando, per esempio, il metodo di Fletcher-Powell [16]. Notiamo che questa procedura riguarda solo il modulo della risposta. Di conseguenza, l'algoritmo di ottimizzazione può fornire per i parametri valori che corrispondono a filtri instabili, cioè i poli e gli zeri di ogni blocco del secondo ordine sono non vincolati. Invece che porre delle condizioni sui parametri, Steiglitz propone di esaminare le radici di ogni fattore del secondo ordine al termine della procedura di minimizzazione e, se un polo (o uno zero) cade al di fuori del cerchio unitario, di sostituirlo con il suo reciproco, così che il modulo della risposta resta invariato. Si è trovato che si può ottenere un'ulteriore riduzione dell'errore continuando l'ottimizzazione a partire dai nuovi valori.

Il seguente esempio [14] illustra l'uso della procedura esposta prima.

³ Il segno di A non appare nella minimizzazione perché viene considerato soltanto l'errore in ampiezza.

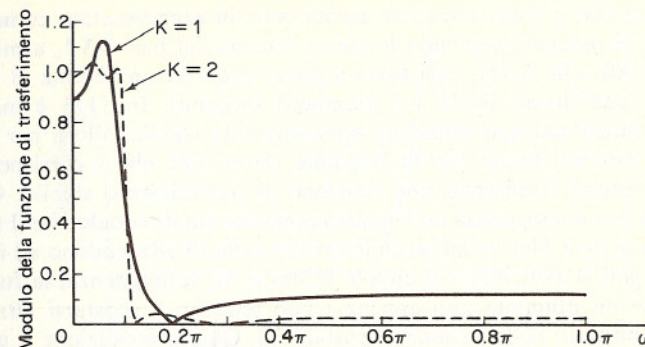


Fig. 5.28 Esempio delle risposte in frequenza ottenute minimizzando l'errore quadratico medio (da Steiglitz [14]).

È richiesto un filtro passa-basso con frequenza di taglio di 0.1π e con valori preassegnati come in fig. 5.27. Cioè

$$|H_d(e^{j\omega})| = \begin{cases} 1, & \omega = 0, 0.01\pi, 0.02\pi, \dots, 0.09\pi \\ 0.5, & \omega = 0.1\pi \\ 0, & \omega = 0.11\pi, 0.12\pi, \dots, 0.19\pi \\ 0, & \omega = 0.2\pi, 0.3\pi, \dots, 0.9\pi, \pi \end{cases}$$

È da notare che non è richiesto che le frequenze fissate $\{\omega_i\}$ siano uniformemente spaziate. La fig. 5.28 mostra i risultati della procedura di ottimizzazione per $K=1$ e $K=2$.

5.3.2 Minimizzazione dell'errore di ordine p

Deczky [17] ha generalizzato in vari modi la procedura descritta nel precedente paragrafo. Invece di minimizzare l'errore quadratico medio, si minimizza una media pesata della p -esima potenza dell'errore e inoltre questa tecnica si applica sia al modulo che al ritardo di gruppo. Di conseguenza, la funzione errore da minimizzare è una approssimazione discreta di

$$E_p = \int_0^\pi W(\omega) [|H(e^{j\omega})| - |H_d(e^{j\omega})|]^p d\omega \quad (5.40)$$

o di

$$E_p = \int_0^\pi W(\omega) [\tau(\omega) - \tau_d(\omega)]^p d\omega \quad (5.41)$$

dove il ritardo di gruppo τ è definito come

$$\tau(\omega) = -\frac{d}{d\omega} \{\arg [H(e^{j\omega})]\} \quad (5.42)$$

Si suppone che il filtro numerico desiderato sia rappresentato come nella (5.38). Se si procede, seguendo lo stesso schema del par. 5.3.1, a minimizzare la (5.40) o la (5.41), ci si trova a dover risolvere un sistema di $4K+1$ equazioni non lineari in $4K+1$ parametri incogniti. In [17] è mostrato che se si inizia con una soluzione approssimante stabile, allora per $p \geq 2$ esiste un minimo locale per la funzione errore tale che i corrispondenti parametri ottimi forniscono una funzione di trasferimento stabile. Questo risultato è una conseguenza dell'osservazione che sia il modulo che il ritardo di gruppo di ogni blocco del secondo ordine nella (5.38) tendono ad infinito quando i poli si avvicinano al circolo unitario. Di conseguenza, la funzione errore diventa illimitata non appena i poli tendono a spostarsi attraverso il circolo unitario nella regione di instabilità. Questa condizione è utile in quanto costituisce una « barriera » al movimento dei poli nel risolvere le equazioni non lineari usando l'algoritmo di Fletcher-Powell [16].

Come esempio di applicazione di questa tecnica, è stato progettato un filtro ellittico passa-basso del quarto ordine. La caratteristica di attenuazione ($-20 \log_{10}|H(e^{j\omega})|$) di questo filtro è mostrata in fig. 5.29(a) ed il ritardo di gruppo in fig. 5.29(b). Poiché il filtro ellittico ha una risposta di ampiezza ottima ma una fase non lineare, si è cercato un equalizzatore passa-tutto da collegare in cascata in modo tale da migliorare la curva di fase, cioè per appiattire la curva del ritardo di gruppo nella banda passante. Modificando così il progetto e usando un valore di $p=10$ nella (5.41), si ottiene la curva del ritardo di gruppo di fig. 5.29(b). Si può notare come tale ritardo di gruppo abbia un andamento ad oscillazione uniforme nella banda passante. Ciò è dovuto al fatto che è stato usato un valore relativamente alto per p . Infatti, può essere mostrato che in generale per $p \rightarrow \infty$, la soluzione ottima tende ad una vera approssimazione ad oscillazione uniforme.

5.3.3 Progetto del filtro inverso con i minimi quadrati

Nelle due procedure descritte precedentemente, il filtro veniva specificato nel dominio della frequenza e il sistema di equazioni risultanti era non lineare nei parametri del filtro. Una procedura alternativa, basata su un'approssimazione con i minimi quadrati dell'inverso del filtro desiderato, conduce ad un sistema di equazioni lineari. In questa tecnica il filtro viene specificato con i primi L campioni della risposta all'impulso desiderata:

$$\{h_d(n)\}, \quad n = 0, 1, \dots, L-1$$

Nella nostra discussione assumeremo che la funzione di trasferimento del filtro sia della forma

$$H(z) = \frac{b_0}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (5.43)$$

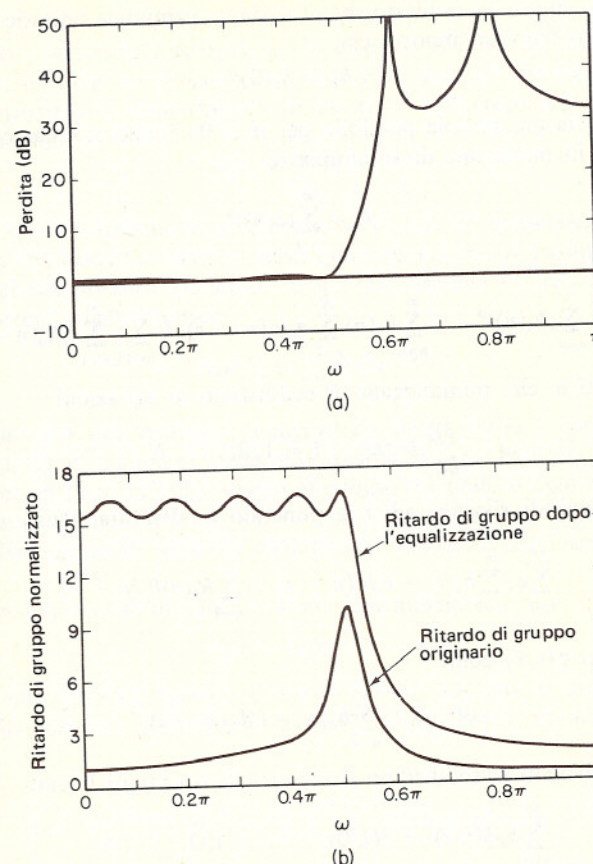


Fig. 5.29 (a) caratteristica di perdita di un filtro ellittico passa-basso del quarto ordine; (b) ritardo di gruppo di un filtro ellittico passa-basso del quarto ordine equalizzato tramite un filtro passa-tutto a quattro sezioni ($p=10$. Da Deczky [17]).

Generalizzazioni di questa procedura che consentono l'introduzione sia di poli che di zeri sono state discusse da Shanks [19] e da Burrus e Parks [18]. Nel modello più semplice da noi adottato, il progetto del filtro è basato sul criterio che l'uscita dell'inversa di $H(z)$ deve approssimare un campione unitario quando l'ingresso è $h_d(n)$. Se con $v(n)$ indichiamo l'uscita del sistema inverso con funzione di trasferimento $1/H(z)$, allora

$$V(z) = \frac{H_d(z)}{H(z)}$$

Di conseguenza possiamo scrivere la formula ricorsiva

$$b_0 v(n) = h_d(n) - \sum_{r=1}^N a_r h_d(n-r) \quad (5.44)$$

Ricordiamo che viene richiesto che $v(n)$ sia un campione unitario. Di conseguenza è ragionevole imporre che

$$b_0 = h_d(0)$$

e che $v(n)$ sia più piccola possibile per $n > 0$. Perciò scegliamo i restanti coefficienti in modo tale da minimizzare

$$E = \sum_{n=1}^{\infty} (v(n))^2$$

Dalla (5.44) si ha

$$E = \frac{1}{b_0^2} \sum_{n=1}^{\infty} (h_d(n))^2 - 2 \sum_{n=1}^{\infty} h_d(n) \sum_{r=1}^N a_r h_d(n-r) + \sum_{n=1}^{\infty} \left[\sum_{r=1}^N a_r h_d(n-r) \right]^2$$

I coefficienti a_i che minimizzano E soddisfano le equazioni

$$\frac{\partial E}{\partial a_i} = 0, \quad i = 1, 2, \dots, N$$

Derivando quindi rispetto ad a_i e ponendo la derivata uguale a zero si ottiene

$$\sum_{r=1}^N a_r \sum_{n=1}^{\infty} h_d(n-r) h_d(n-i) = \sum_{n=1}^{\infty} h_d(n) h_d(n-i)$$

Se definiamo $\phi(i, r)$ come

$$\phi(i, r) = \sum_{n=1}^{\infty} h_d(n-r) h_d(n-i)$$

allora i coefficienti a_i soddisfano il sistema di equazioni lineari

$$\sum_{r=1}^N a_r \phi(i, r) = \phi(i, 0), \quad i = 1, 2, \dots, N \quad (5.45)$$

Queste equazioni possono essere risolte utilizzando una qualsiasi tecnica convenzionale. Può essere mostrato [18, 19] che la matrice delle quantità $\phi(i, r)$ è definita positiva. Di conseguenza una procedura particolarmente efficiente è quella fornita da Levinson [20].

5.4 PROPRIETÀ DEI FILTRI NUMERICI FIR

Nei paragrafi precedenti sono state illustrate delle tecniche di progetto relative ai filtri numerici con risposta all'impulso di durata infinita. Benché tali filtri abbiano delle caratteristiche molto attraenti, essi hanno anche un certo numero di svantaggi. Per esempio, se si desidera sfruttare il vantaggio offerto dalla velocità di calcolo della FFT e realizzare un filtro come è stato discusso nel cap. 3, è essenziale avere una risposta all'impulso di durata finita. Inoltre, gli esempi dei precedenti paragrafi mettono in chiaro il fatto che i filtri IIR consentono di ottenere generalmente delle eccellenti

risposte di ampiezza con l'inconveniente però di risposte di fase non lineari. Al contrario, i filtri FIR possono avere fase esattamente lineare. Di conseguenza, le tecniche di progetto dei filtri FIR sono di interesse notevole.

La funzione di trasferimento di un filtro FIR causale è della forma

$$H(z) = \sum_{n=0}^{N-1} h(n) z^{-n}$$

cioè $H(z)$ è un polinomio in z^{-1} di grado $N-1$. Di conseguenza $H(z)$ ha $N-1$ zeri che possono essere disposti ovunque nel piano z al finito e $N-1$ poli che giacciono tutti nel punto $z=0$. La risposta in frequenza $H(e^{j\omega})$ è il polinomio trigonometrico

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n) e^{-j\omega n} \quad (5.46)$$

Ricordiamo che una qualsiasi sequenza di durata finita è completamente specificata da N campioni della sua trasformata di Fourier, e quindi il progetto di un filtro FIR può essere effettuato trovando o i coefficienti della sua risposta all'impulso, oppure N campioni della sua risposta in frequenza. Nei paragrafi seguenti discuteremo esempi relativi ad entrambi i metodi.

Se la risposta all'impulso soddisfa la condizione

$$h(n) = h(N-1-n) \quad (5.47)$$

allora il filtro ha fase lineare. Come si è visto nel cap. 4, ciò può essere facilmente mostrato sostituendo la (5.47) nella (5.46) ottenendo di conseguenza

$$H(e^{j\omega}) = \begin{cases} e^{-j\omega((N-1)/2)} \left[h\left(\frac{N-1}{2}\right) + \sum_{n=0}^{(N-3)/2} 2h(n) \cos\left(\omega\left(n - \frac{N-1}{2}\right)\right) \right], & N \text{ dispari} \quad (5.48a) \\ e^{-j\omega((N-1)/2)} \left[\sum_{n=0}^{N/2-1} 2h(n) \cos\left(\omega\left(n - \frac{N-1}{2}\right)\right) \right], & N \text{ pari} \quad (5.48b) \end{cases}$$

Si può vedere da queste espressioni che la condizione (5.47) implica uno sfasamento lineare corrispondente ad un ritardo di $(N-1)/2$ campioni. Notiamo che per il caso in cui N è dispari, lo sfasamento corrisponde a un ritardo di un numero intero di campioni, mentre per N pari il ritardo è un numero intero più un mezzo.⁴ Questa distinzione tra valori di N pari e dispari è spesso di considerevole importanza nel progetto e nella realizzazione di filtri FIR. Esempi di risposte all'impulso aventi fase lineare sono mostrati in fig. 5.30.

⁴ Può essere mostrato [6] che l'eq. (5.47) è sia necessaria che sufficiente affinché un filtro numerico causale abbia una fase lineare. Di conseguenza soltanto i filtri FIR possono avere una fase lineare.

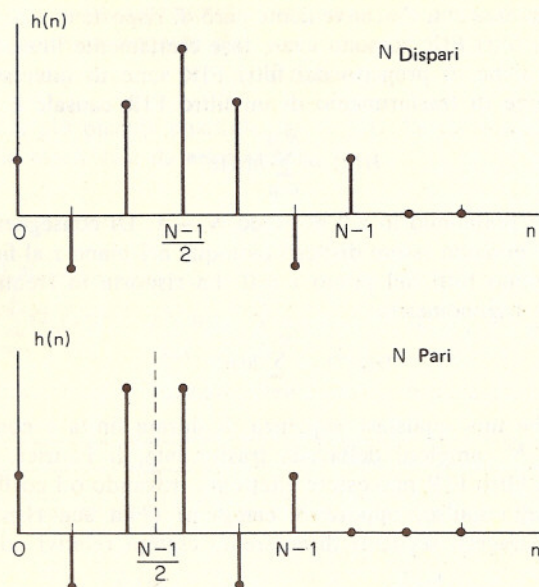


Fig. 5.30 Risposte all'impulso tipiche per filtri FIR a fase lineare.

Dedicheremo quasi interamente la discussione seguente ai filtri a fase lineare, in quanto quest'ultima, oltre a essere utile e talora necessaria, semplifica spesso i procedimenti di progetto.

5.5 PROGETTO DI FILTRI FIR CON L'USO DI FINESTRE

L'approccio più immediato al progetto di filtri FIR consiste nell'ottenere una risposta all'impulso di lunghezza finita troncando una risposta di durata infinita. Se indichiamo con $H_d(e^{j\omega})$ la risposta in frequenza ideale desiderata, si può scrivere

$$H_d(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h_d(n) e^{-j\omega n} \quad (5.49a)$$

dove $h_d(n)$ è la sequenza che rappresenta la risposta all'impulso corrispondente, cioè

$$h_d(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_d(e^{j\omega}) e^{j\omega n} d\omega \quad (5.49b)$$

In generale, la $H_d(e^{j\omega})$ per un filtro selettivo in frequenza può essere costante a tratti, con discontinuità in corrispondenza delle transizioni tra una banda e l'altra. In questi casi la sequenza $h_d(n)$ è di durata infinita e deve essere troncata per ottenere una risposta all'impulso di durata finita. Come abbiamo già notato, le relazioni (5.49) possono essere viste come una rap-

presentazione mediante la serie di Fourier della risposta in frequenza, periodica, $H_d(e^{j\omega})$, dove la sequenza $h_d(n)$ gioca il ruolo dei « coefficienti di Fourier ». Quindi l'approssimazione delle specifiche di un filtro ideale mediante troncamento della risposta all'impulso ideale coincide con lo studio della convergenza delle serie di Fourier, studio che ha avuto un gran numero di contributi sin dalla metà del secolo diciottesimo. Il concetto più noto legato a questa teoria è il *fenomeno di Gibbs*. Nella discussione che segue vedremo come si manifesta questo fenomeno di convergenza non uniforme nel progetto di filtri FIR.

Se $h_d(n)$ ha durata infinita, un modo di ottenere una risposta all'impulso causale e di durata finita è quello di troncare semplicemente $h_d(n)$, cioè definire

$$h(n) = \begin{cases} h_d(n), & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (5.50)$$

In generale, $h(n)$ può essere rappresentata come il prodotto della risposta desiderata e di una « finestra » $w(n)$ di durata finita, cioè

$$h(n) = h_d(n)w(n) \quad (5.51)$$

dove per l'esempio della (5.50) è

$$w(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases} \quad (5.52)$$

Usando il teorema della convoluzione complessa derivato nel cap. 2 si vede che

$$H(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_d(e^{j\theta}) W(e^{j(\omega-\theta)}) d\theta \quad (5.53)$$

In altri termini, $H(e^{j\omega})$ è la convoluzione continua periodica della risposta in frequenza desiderata con la trasformata di Fourier della finestra. Perciò la risposta in frequenza $H(e^{j\omega})$ sarà una versione « smussata » della risposta desiderata $H_d(e^{j\omega})$. La fig. 5.31(a) mostra alcuni tipici esempi delle funzioni $H_d(e^{j\theta})$ e $W(e^{j(\omega-\theta)})$ con riferimento alla relazione (5.53) (entrambe le funzioni sono rappresentate come reali solo per comodità).

Dalla (5.53) si vede che, se $W(e^{j\omega})$ è a banda stretta rispetto alle variazioni di $H_d(e^{j\omega})$, allora $H(e^{j\omega})$ « assomiglia » ad $H_d(e^{j\omega})$. Quindi i criteri di scelta della finestra sono, da un lato, l'esigenza di avere $w(n)$ la più corta possibile in quanto a durata, onde ridurre la complessità dei calcoli nelle operazioni di filtraggio, e dall'altro quella che $W(e^{j\omega})$ sia la più stretta possibile in frequenza, per riprodurre fedelmente la risposta desiderata. Queste esigenze sono contrastanti, come si può vedere nel caso della finestra rettangolare espressa dalla (5.52), dove

$$\begin{aligned} W(e^{j\omega}) &= \sum_{n=0}^{N-1} e^{-j\omega n} = \frac{1 - e^{-j\omega N}}{1 - e^{-j\omega}} \\ &= e^{-j\omega((N-1)/2)} \frac{\sin(\omega N/2)}{\sin(\omega/2)} \end{aligned} \quad (5.54)$$

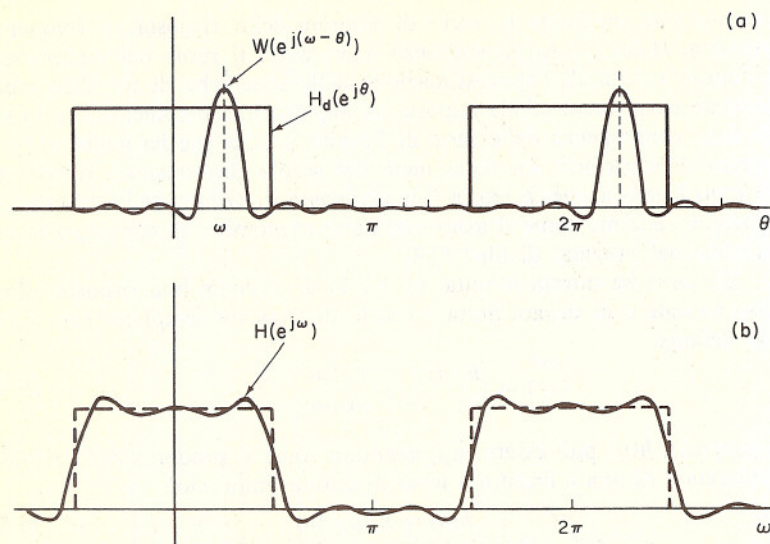


Fig. 5.31 (a) Operazione di convoluzione corrispondente al troncamento della risposta all'impulso desiderata; (b) tipica risposta approssimata che risulta dall'applicazione della finestra alla risposta all'impulso desiderata.

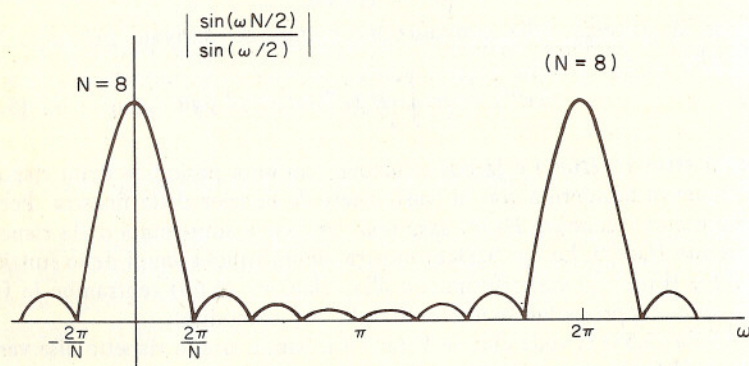


Fig. 5.32 Modulo della trasformata di Fourier di una finestra rettangolare ($N=8$).

Il modulo di $W(e^{j\omega})$ è rappresentato in fig. 5.32 per $N=8$ e la fase è lineare, come si può vedere dall'espressione (5.54). Quando N aumenta, la larghezza del « lobo principale » diminuisce (il lobo principale è definito in maniera arbitraria come la regione compresa tra $\omega = -2\pi/N$ e $\omega = +2\pi/N$).

Nel caso della finestra rettangolare, però, i « lobi laterali » non sono trascurabili ed in effetti, quando N cresce, i valori di picco del lobo principale e dei lobi laterali aumentano in modo tale che l'area sotto ogni lobo resta costante, mentre la larghezza di ogni lobo diminuisce con N . Il risul-

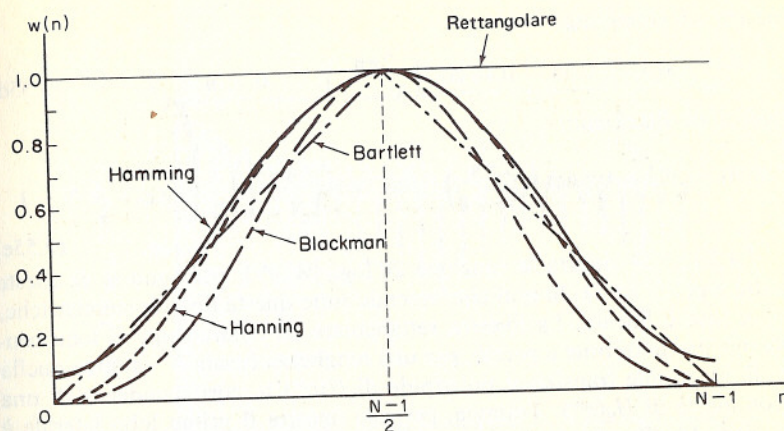


Fig. 5.33 Finestre comunemente usate per il progetto di filtri FIR.

tato di questo fatto è che quando $W(e^{j(\omega-\theta)})$ incontra una discontinuità di $H_d(e^{j\omega})$ con ω crescente, l'integrale di $W(e^{j(\omega-\theta)})H_d(e^{j\theta})$ varia in modo oscillatorio mentre i singoli lobi di $W(e^{j(\omega-\theta)})$ si sovrappongono alla discontinuità. Questo effetto è rappresentato nella fig. 5.31(b). Poiché l'area sotto ogni lobo rimane costante mentre N cresce, ne segue che le oscillazioni hanno semplicemente una frequenza maggiore ma non diminuiscono in ampiezza, quando N aumenta. È ben noto nella teoria della serie di Fourier che questa convergenza non uniforme, il fenomeno di Gibbs, può essere ridotto usando un troncamento meno brusco della serie di Fourier.

Si può diminuire l'altezza dei lobi laterali facendo tendere a zero dolcemente la finestra ai suoi estremi; si paga però questo miglioramento con un allargamento del lobo principale e quindi con una transizione meno ripida in corrispondenza della discontinuità. La fig. 5.33 mostra alcuni esempi di finestre di uso comune. Queste finestre sono definite dalle relazioni seguenti [21]:

Finestra rettangolare: $w(n) = 1, \quad 0 \leq n \leq N-1$ (5.55a)

Finestra di Bartlett: $w(n) = \begin{cases} \frac{2n}{N-1}, & 0 \leq n \leq \frac{N-1}{2} \\ 2 - \frac{2n}{N-1}, & \frac{N-1}{2} \leq n \leq N-1 \end{cases}$ (5.55b)

Finestra di Hanning: $w(n) = \frac{1}{2} \left[1 - \cos \left(\frac{2\pi n}{N-1} \right) \right], \quad 0 \leq n \leq N-1$ (5.55c)

Finestra di Hamming:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (5.55d)$$

Finestra di Blackman:

$$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (5.55e)$$

La fig. 5.34 riporta la funzione $20 \log_{10} |W(e^{j\omega})|$ per ognuna di queste finestre e per $N = 51$. Si noti che, essendo tutte queste finestre simmetriche, la loro fase è lineare. La finestra rettangolare ha chiaramente il lobo centrale più stretto di tutte e perciò, per una lunghezza fissata N , sarebbe quella che dà luogo alle transizioni più ripide di $H(e^{j\omega})$ in corrispondenza di una discontinuità di $H_d(e^{j\omega})$. Tuttavia, per tale finestra il primo lobo laterale è solo circa 13 dB sotto il picco principale, per cui si hanno notevoli oscillazioni di $H(e^{j\omega})$ in corrispondenza di una discontinuità di $H_d(e^{j\omega})$. Se si raccorda in maniera dolce la finestra verso lo zero, si riducono molto i lobi laterali; è chiaro però che si paga il prezzo di un lobo principale molto più largo con conseguenti transizioni meno marcate alle discontinuità di $H_d(e^{j\omega})$.

Kaiser [4] ha proposto una famiglia flessibile di finestre definite come

$$w(n) = \frac{I_0\left[\omega_a \sqrt{\left(\frac{N-1}{2}\right)^2 - \left[n - \left(\frac{N-1}{2}\right)\right]^2}\right]}{I_0\left[\omega_a \left(\frac{N-1}{2}\right)\right]}$$

dove $I_0(\cdot)$ è la funzione di Bessel modificata di ordine zero del primo tipo. Kaiser ha dimostrato che queste finestre sono quasi ottime, nel senso che hanno la massima energia nel lobo principale per un fissato valore di picco dei lobi laterali. Il parametro ω_a può essere variato in modo da conciliare larghezza del lobo principale e altezza dei lobi secondari. Valori tipici di $\omega_a((N-1)/2)$ sono nel campo $4 < \omega_a((N-1)/2) < 9$.

Per illustrare l'uso delle finestre nel progetto dei filtri, consideriamo il progetto di un filtro passa-basso. Tenendo presente fin d'ora la necessità di un ritardo per ottenere un filtro causale a fase lineare, definiamo la risposta in frequenza desiderata come

$$H_d(e^{j\omega}) = \begin{cases} e^{-j\omega\alpha}, & |\omega| \leq \omega_c \\ 0, & \text{altrove} \end{cases}$$

La risposta all'impulso corrispondente è

$$h_d(n) = \begin{cases} \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} e^{j\omega(n-\alpha)} d\omega \\ \frac{\sin[\omega_c(n-\alpha)]}{\pi(n-\alpha)}, & n \neq \alpha \end{cases}$$

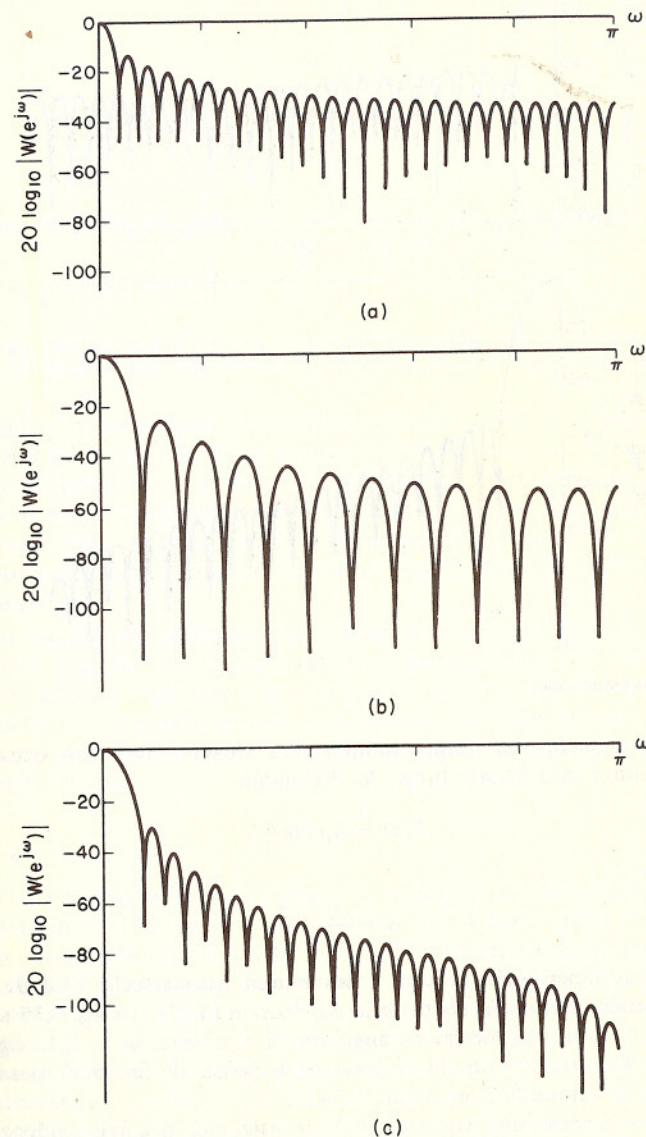


Fig. 5.34 Trasformate di Fourier delle finestre della fig. 5.33: (a) rettangolare; (b) Bartlett (triangolare); (c) Hanning; (d) Hamming; (e) Blackman.

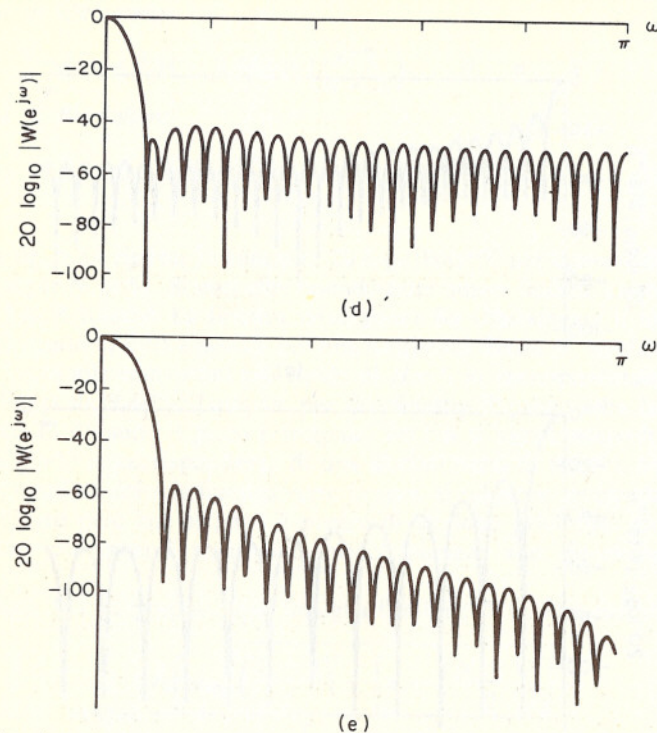


Fig. 5.34 (continuazione)

Chiaramente, $h_d(n)$ ha durata infinita. Per ricavare un filtro causale di durata finita a fase lineare lungo N , definiamo

$$h(n) = h_d(n)w(n)$$

dove

$$\alpha = \frac{N-1}{2}$$

Si verifica facilmente che se $w(n)$ è simmetrica, questa scelta di α dà luogo ad una sequenza $h(n)$ che soddisfa la condizione (5.47). La fig. 5.35 mostra un grafico di $h_d(n)$ per finestra rettangolare, $N = 51$ e $\omega_c = \pi/2$. La fig. 5.36 mostra $20 \log_{10} |H(e^{j\omega})|$ per la risposta all'impulso di fig. 5.35 pesata da ognuna delle cinque finestre di fig. 5.34.

Si noti la crescente larghezza della transizione, in corrispondenza dell'aumento di estensione del lobo principale, e l'attenuazione crescente nella banda oscura, corrispondente a minori ampiezze dei lobi laterali.

Dalla relazione (5.54) si nota che la larghezza del lobo principale è inversamente proporzionale a N . Questo è vero in generale ed è illustrato per una finestra di Hamming nella fig. 5.37, da cui appare chiaramente

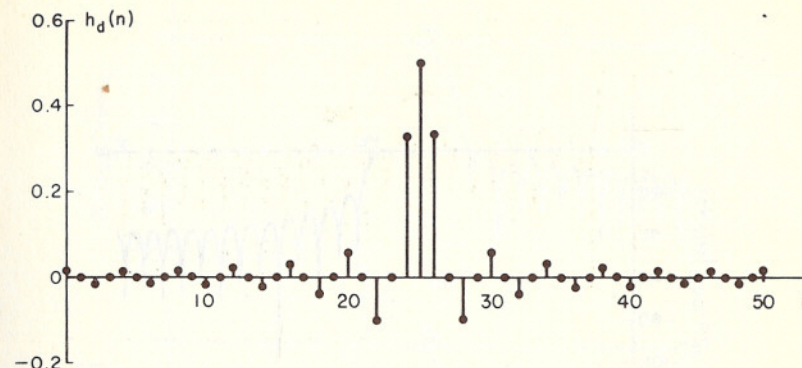


Fig. 5.35 Risposta all'impulso troncata di un filtro passa-basso ideale. (Il ritardo è di 25 campioni, la lunghezza totale è 51 campioni e la frequenza di taglio è $\omega_c = \pi/2$).

che quando N raddoppia, l'estensione del lobo centrale si dimezza. La fig. 5.38 illustra l'effetto di un aumento di N sulla regione di transizione nel caso di progetto di un filtro passa-basso. È chiaro che l'attenuazione minima in banda oscura rimane essenzialmente costante, in quanto dipende dalla forma della finestra, mentre l'ampiezza della regione di transizione in corrispondenza di una discontinuità di $H_d(e^{j\omega})$ dipende dalla lunghezza della finestra.

Gli esempi che abbiamo dato illustrano i principi generali del metodo della pesatura con finestre per il progetto di filtri FIR. Il procedimento di progetto può essere controllato in qualche misura attraverso la scelta della forma e della durata della finestra. Ad esempio, per una data attenuazione in banda oscura, è vero in generale che N soddisfa un'equazione della forma

$$N = \frac{A}{\Delta\omega}$$

dove $\Delta\omega$ è la larghezza della transizione [all'incirca la larghezza del lobo principale di $W(e^{j\omega})$] ed A è una costante che dipende dalla forma della finestra. Come abbiamo visto, la forma della finestra è fondamentale per quanto riguarda l'attenuazione minima in banda oscura. Per le finestre che abbiamo presentato, i parametri di base per il progetto di filtri passa-basso sono riassunti in tab. 5.2. Va notato che i valori della tabella sono approssimati; essi dipendono in certa misura da N e dalla frequenza di taglio del filtro desiderato. Le finestre di Kaiser hanno un parametro variabile, ω_a , la cui scelta controlla il rapporto tra altezza dei lobi laterali e larghezza del lobo principale. Tabelle e curve relative all'uso di queste finestre sono fornite da Kaiser in [4,22].

I principi fondamentali illustrati dai nostri esempi sono validi in generale e possono essere applicati al progetto di qualsiasi filtro definibile mediante una risposta in frequenza desiderata. In questo senso, la tecnica

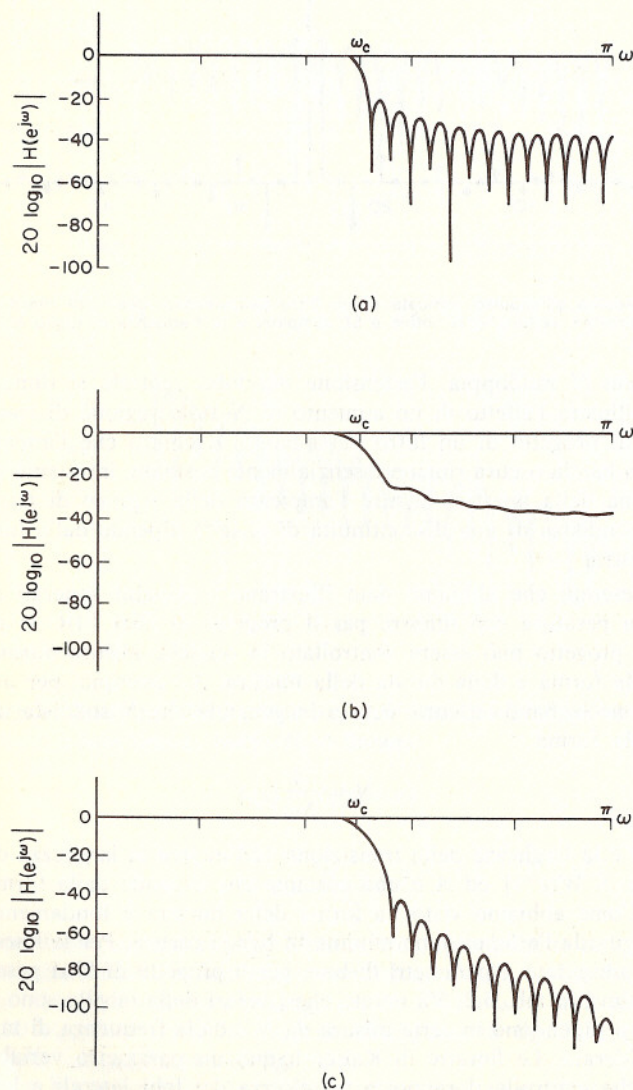


Fig. 5.36 Effetto delle diverse finestre per l'esempio di fig. 5.35: (a) rettangolare; (b) Bartlett; (c) Hanning; (d) Hamming; (e) Blackman.

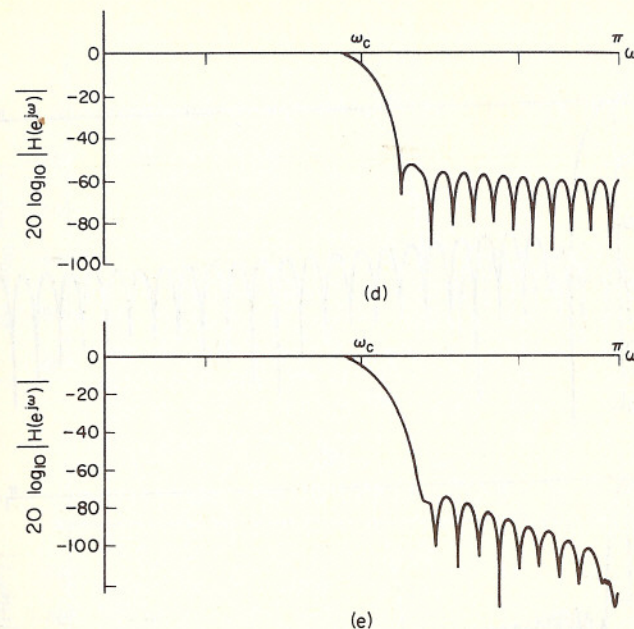


Fig. 5.36 (continuazione)

ha la caratteristica di una notevole generalità. Esiste però la difficoltà di calcolare l'integrale della (5.49 b). Se $H_d(e^{j\omega})$ non è esprimibile in termini di funzioni semplici per cui si possa eseguire l'integrazione, occorre ottenere un'approssimazione di $h_d(n)$ campionando $H_d(e^{j\omega})$ e usando la trasformata di Fourier discreta inversa per calcolare

$$\begin{aligned}\tilde{h}_d(n) &= \frac{1}{M} \sum_{k=0}^{M-1} H_d(e^{j(2\pi/M)k}) e^{j(2\pi/N)kn} \\ &= \sum_{r=-\infty}^{\infty} h_d(n + rM)\end{aligned}$$

Se M è grande, $\tilde{h}_d(n)$ può essere ritenuta una buona approssimazione di $h_d(n)$ nell'intervallo ricoperto dalla finestra. Un'altra limitazione di questo

Tab. 5.2

Finestra	Altezza massima dei lobi laterali (dB)	Larghezza del lobo principale	Attenuazione minima in banda oscura
Rettangolare	-13	$4\pi/N$	-21
Bartlett	-25	$8\pi/N$	-25
Hanning	-31	$8\pi/N$	-44
Hamming	-41	$8\pi/N$	-53
Blackman	-57	$12\pi/N$	-74

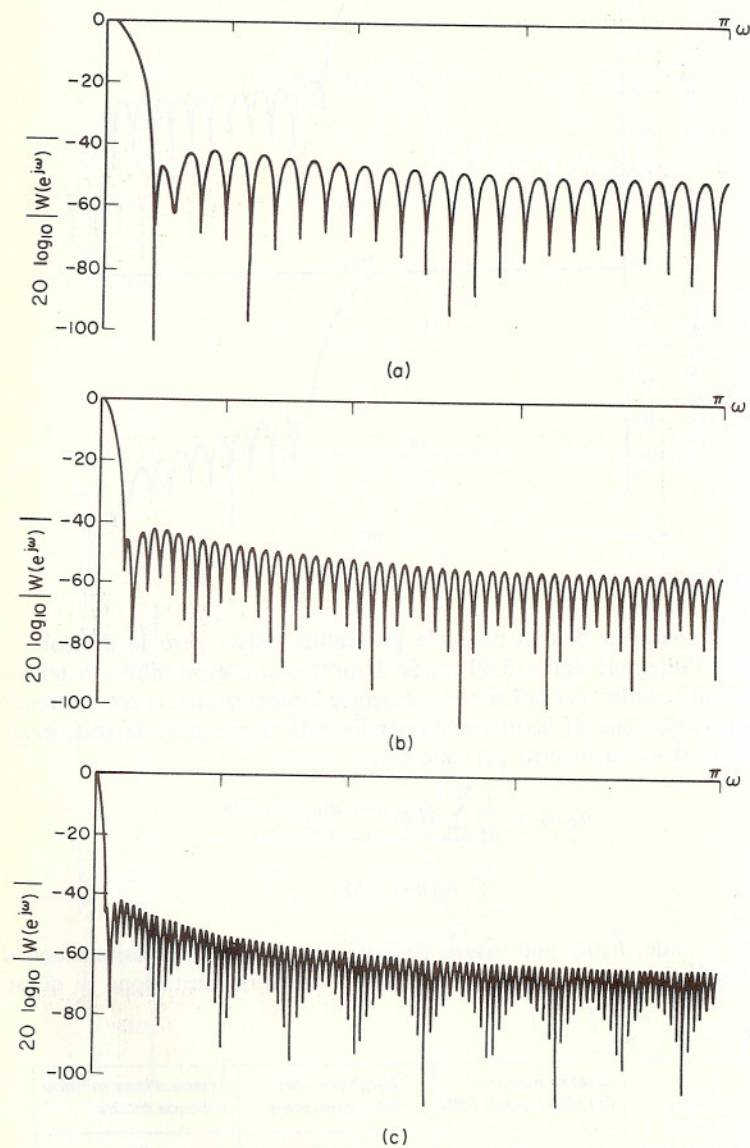


Fig. 5.37 Dipendenza della trasformata di Fourier della finestra di Hamming dalla lunghezza della finestra: (a) $N=51$; (b) $N=101$; (c) $N=201$.

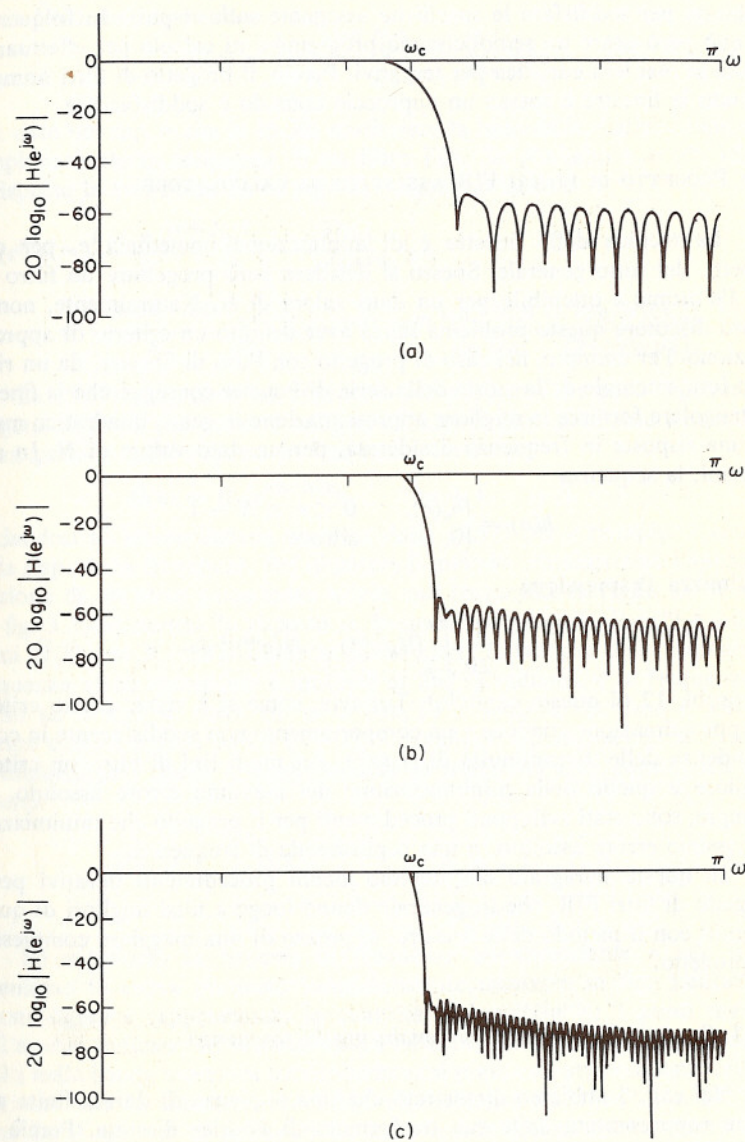


Fig. 5.38 Effetto della lunghezza della finestra nel progetto di un filtro (passa-basso, $\omega_c=\pi/2$ e finestra di Hamming): (a) $N=51$; (b) $N=101$; (c) $N=201$.

metodo è la difficoltà di determinare a priori il tipo di finestra e la durata N richieste per soddisfare le specifiche assegnate sulla risposta in frequenza.⁵ Si può però usare un semplicissimo programma di calcolo per effettuare le scelte in maniera euristica per tentativi. Perciò, il progetto di filtri numerici usando le finestre è spesso un approccio comodo e soddisfacente.

5.6 PROGETTO DI FILTRI FIR ASSISTITO DA CALCOLATORE

La tecnica delle finestre è di applicazione immediata e, per certi aspetti, del tutto generale. Spesso si desidera però progettare un filtro che sia l'« ottimo » ottenibile per un dato valore di N . Naturalmente, non ha senso discutere questo problema senza aver definito un criterio di approssimazione. Per esempio, nel caso di progetto con l'uso di finestre, da un risultato fondamentale della teoria della serie di Fourier consegue che la finestra rettangolare fornisce la migliore approssimazione in senso quadratico medio di una risposta in frequenza desiderata, per un dato valore di N . In altri termini, la sequenza

$$h(n) = \begin{cases} h_d(n), & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

minimizza l'espressione

$$\varepsilon^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_d(e^{j\omega}) - H(e^{j\omega})|^2 d\omega$$

(v. probl. 12 di questo capitolo). Tuttavia, come si è visto, questo criterio di approssimazione conduce a un comportamento non soddisfacente in corrispondenza delle discontinuità di $H_d(e^{j\omega})$. Per molti tipi di filtri, un criterio migliore è quello della minimizzazione del massimo errore assoluto. Ad esempio, sono stati sviluppati procedimenti per il progetto che minimizzano il massimo errore assoluto in una o più bande di frequenza.

In questo paragrafo discuteremo alcuni procedimenti iterativi per il progetto di filtri FIR, che in generale danno luogo a filtri migliori di quelli ottenuti con il metodo delle finestre, al prezzo di una maggiore complessità di progetto.

5.6.1 Progetto basato sul campionamento in frequenza

Nel cap. 3 abbiamo dimostrato che una sequenza di durata finita può essere rappresentata dalla sua trasformata di Fourier discreta. Perciò un filtro FIR ha una rappresentazione in termini dei « campioni frequenziali » del tipo

$$\tilde{H}(k) = H(z)|_{z=e^{j(2\pi/N)k}} = \sum_{n=0}^{N-1} h(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

⁵ In [22] è esposta una procedura sistematica relativa alle finestre di Kaiser.

Come mostrato nel cap. 3, $H(z)$ può essere rappresentata in termini dei valori $\tilde{H}(k)$ dall'espressione

$$H(z) = \frac{1 - z^{-N}}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - e^{j(2\pi/N)k} z^{-1}} \quad (5.56)$$

Si è visto nel cap. 4 che la (5.56) costituisce la base della realizzazione con campionamento in frequenza di un filtro FIR. Se poniamo $z = e^{j\omega}$, allora la risposta in frequenza ha la seguente rappresentazione

$$\begin{aligned} H(e^{j\omega}) &= \frac{1 - e^{-j\omega N}}{N} \sum_{k=0}^{N-1} \frac{\tilde{H}(k)}{1 - e^{j(2\pi/N)k} e^{-j\omega}} \\ &= \frac{e^{-j\omega(N-1)/2}}{N} \sum_{k=0}^{N-1} \tilde{H}(k) e^{j\pi k(1-1/N)} \frac{\sin [N(\omega - (2\pi/N)k)/2]}{\sin [(\omega - (2\pi/N)k)/2]} \end{aligned} \quad (5.57)$$

La (5.57) suggerisce un approccio semplice, ma un poco ingenuo, al progetto di filtri, che consiste nello specificare il filtro mediante i campioni di un periodo della risposta in frequenza desiderata

$$\tilde{H}(k) = H_d(e^{j(2\pi/N)k}), \quad k = 0, 1, \dots, N-1$$

affidandosi all'interpolazione espressa dalla (5.57) per « riempire i vuoti » nella risposta in frequenza. Per illustrare il metodo, consideriamo l'approssimazione di un filtro passa-basso ideale con frequenza di taglio $\omega_c = \pi/2$. La fig. 5.39(a) mostra la risposta in frequenza desiderata $H_d(e^{j\omega})$ e i campioni $\tilde{H}(k)$ per $N = 33$. Come si può vedere, il modulo della risposta in frequenza viene specificato a multipli di $2\pi/33$ radianti, e la frequenza di taglio $\omega_c = \pi/2$ si trova tra $\omega = 16\pi/33$ e $18\pi/33$. La fase è assunta lineare con un ritardo pari a $(N-1)/2$ campioni. Naturalmente, la risposta all'impulso può essere ottenuta usando la trasformata di Fourier discreta inversa:

$$\begin{aligned} h(n) &= \frac{1}{N} \sum_{k=0}^{N-1} \tilde{H}(k) e^{j(2\pi/N)kn} \quad n = 0, 1, \dots, N-1 \\ &= 0 \quad \text{altrove} \end{aligned} \quad (5.58)$$

Se valutiamo la risposta in frequenza corrispondente a tale filtro, otteniamo la curva piuttosto insoddisfacente mostrata in fig. 5.40(a). In questa figura è rappresentata la quantità $20 \log_{10} |H(e^{j\omega})|$: i punti marcati sull'asse ω indicano i campioni fissati nella banda passante, mentre quelli scelti nella banda attenuata corrispondono ai punti con attenuazione infinita. Notiamo che la transizione avviene rapidamente tra $16\pi/33$ e $18\pi/33$; tuttavia, il minimo di attenuazione nella banda oscura risulta essere inferiore a 20 dB. Questi filtri sarebbero insoddisfacenti per molti scopi. Come abbiamo ripetutamente visto, un modo per migliorare l'attenuazione nella banda oscura consiste nell'ampliare la banda di transizione. In questo caso, ciò può essere fatto facilmente permettendo ad un campione che si trova al limite tra la banda passante e la banda attenuata di assumere un valore differente da 0 oppure da 1, come è mostrato in fig. 5.39(b). La

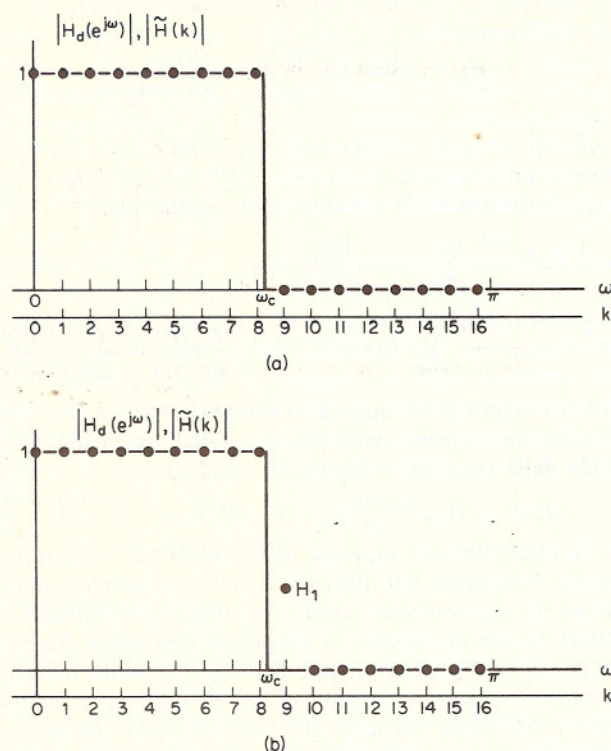


Fig. 5.39 Scelta dei campioni della risposta in frequenza di un filtro passa-basso ideale: (a) senza campioni di transizione; (b) con un campione di transizione H_1 .

figura 5.40(b) mostra la risposta in frequenza per $H_1 = 0.5$. Notiamo che la banda di transizione è ora ampia circa il doppio, ma il minimo di attenuazione nella banda oscura è ora considerevolmente aumentato.

La (5.57) mostra che $H(e^{j\omega})$ è una funzione lineare dei parametri $\tilde{H}(k)$. Di conseguenza si possono impiegare tecniche di ottimizzazione lineare per variare tali parametri allo scopo di ottenere la miglior approssimazione al filtro desiderato. Questo approccio, proposto per la prima volta da Gold e Jordan [23], e sviluppato da Rabiner ed altri [24], è stato usato per progettare una varietà di filtri.

Per esempio, nel caso che stiamo discutendo, si potrà usare una semplice tecnica del gradiente per cercare il valore di H_1 che rende minimo l'errore massimo nella banda passante o nella banda attenuata. La figura 5.40(c) mostra la risposta per $H_1 = 0.3904$, che è il valore che minimizza l'errore (massimizza l'attenuazione) nella banda oscura

$$\frac{20\pi}{33} \leq |\omega| \leq \pi$$

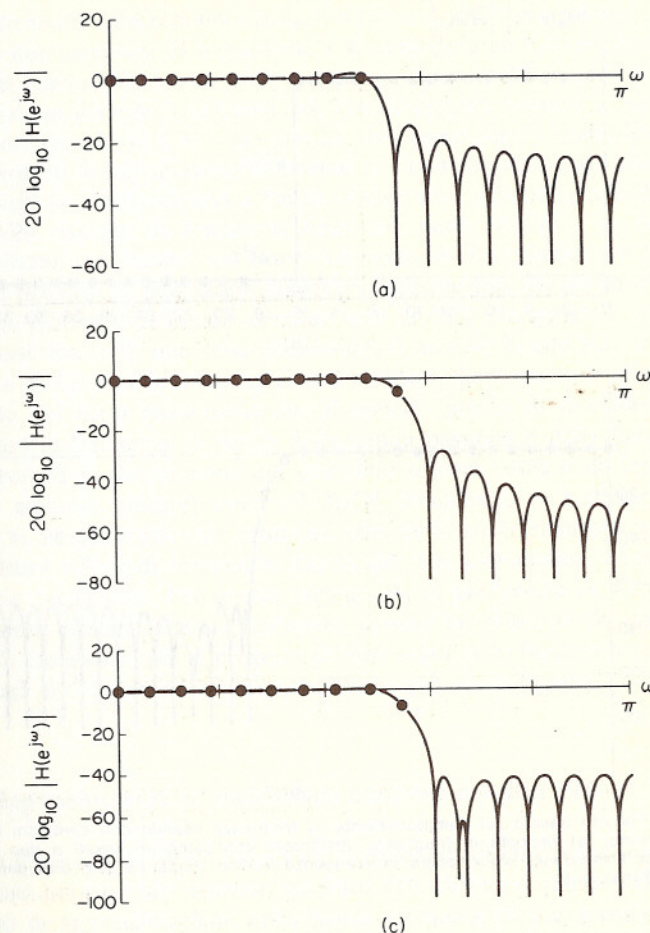


Fig. 5.40 Effetto di un singolo campione di transizione: (a) $H_1 = 0$ (nessun campione di transizione); (b) $H_1 = 0.5$; (c) $H_1 = 0.3904$.

Risulta evidente che l'attenuazione nella banda oscura è significativamente migliorata. Se sono richiesti ulteriori miglioramenti, si può ampliare ancora la banda di transizione permettendo ad un secondo⁶ campione di differire da 1 o da 0. Se N è mantenuto fisso, ciò dà luogo a una regione di transizione larga il doppio, consentendo però una maggiore attenuazione. Ovviamente, se raddoppiamo N , si possono lasciar variare due campioni nella zona di transizione senza che se ne alteri la larghezza. La fig. 5.41(a) mo-

⁶ Rabiner ed altri [24] forniscono risultati relativi a filtri passa-basso aventi fino a quattro valori di transizione variabili.

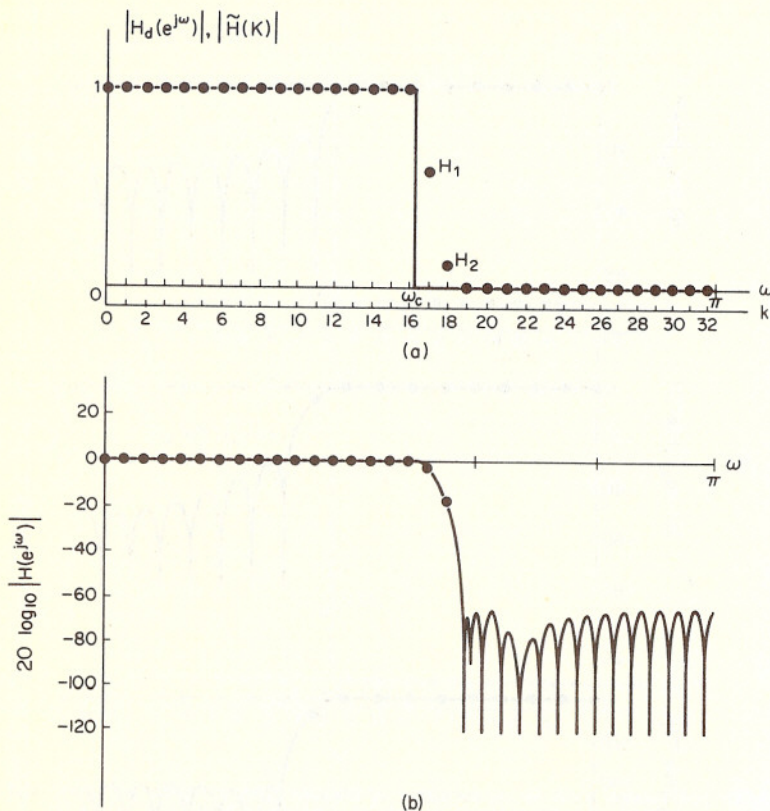


Fig. 5.41 Progetto basato sul campionamento in frequenza usando due campioni di transizione: (a) risposta in frequenza desiderata con campioni fissati e due campioni di transizione; (b) risposta in frequenza ottima risultante per due campioni di transizione.

stra tale insieme di campioni per l'esempio che abbiamo discusso per $N = 65$.⁷ La figura 5.41(b) mostra $20 \log_{10} |H(e^{j\omega})|$ per $N = 65$ e

$$H_1 = \tilde{H}(17) = H(e^{j(34\pi/65)}) = 0.5886$$

$$H_2 = \tilde{H}(18) = H(e^{j(36\pi/65)}) = 0.1065$$

Tali valori sono molto vicini ai campioni di transizione ottimi che minimizzano l'errore massimo assoluto (massimizzano l'attenuazione) nella

⁷ Si noti che $2 \cdot 33 = 66$ è un numero pari e perciò richiederebbe un ritardo corrispondente a un numero non intero di campioni per la linearità di fase. Anche se si possono ottenere progetti basati sul campionamento in frequenza per N pari con difficoltà non maggiore che per N dispari, esistono spesso ragioni pratiche per scegliere N dispari [25].

banda attenuata. Confrontando le fig. 5.41(b) e 5.40(c), si vede che, quando si usano due campioni di transizione e si raddoppia all'incirca N (da 33 a 65), l'attenuazione in banda oscura cresce di circa 24 dB, mentre la banda di transizione diventa lievemente più stretta ($6\pi/65$ rispetto a $8\pi/66$) di quella corrispondente a $N = 33$ con un solo campione di transizione.

I progetti basati sul campionamento in frequenza sono particolarmente convenienti per filtri selettivi a banda stretta dove soltanto pochi dei campioni della risposta in frequenza sono non nulli [25, 26]. In tali casi, una realizzazione basata sul campionamento in frequenza, come quelle descritte nel cap. 4, può essere considerevolmente più efficiente della convoluzione diretta o della convoluzione con la DFT. In generale, anche se i campioni non nulli non sono pochissimi, il metodo basato sul campionamento in frequenza fornisce eccellenti risultati. Comunque, è chiaro dall'esempio del filtro passa-basso che il metodo manca di flessibilità nello specificare le frequenze di taglio della banda passante e della banda attenuata, poiché la disposizione dei campioni unitari, nulli e di transizione avviene secondo multipli interi di $2\pi/N$. Prendendo N sufficientemente grande, si possono ottenere campioni che sono arbitrariamente vicini ad un qualsiasi valore di frequenza specificato; ma, ovviamente, questo è un approccio inefficiente. Per questa ragione, ed in particolare se il filtro non deve essere necessariamente realizzato usando la struttura del campionamento in frequenza, per il progetto di filtri selettivi in frequenza sono stati sviluppati altri algoritmi con caratteristiche più soddisfacenti.

5.6.2 Approssimazioni ad oscillazione uniforme per filtri FIR

La tecnica di progetto basata sul campionamento in frequenza usa una procedura iterativa per ottenere un filtro FIR che ha il più piccolo errore massimo di approssimazione nella banda attenuata (il più grande minimo di attenuazione) per una durata N fissata, per un insieme di campioni in frequenza prescritti e per un dato insieme di campioni in frequenza variabili. Nel caso di filtri selettivi progettati usando questa tecnica, si ha una limitazione indesiderata sulla scelta delle frequenze di taglio. Inoltre l'errore di approssimazione tende ad essere più alto attorno alla zona di transizione e minore lontano dai campioni di transizione. Sembra essere intuitivamente ragionevole il fatto che, se l'errore di approssimazione fosse distribuito uniformemente in frequenza, una data specifica di progetto dovrebbe essere soddisfatta con un filtro di ordine inferiore di quello che si avrebbe nel caso in cui il filtro soddisfa esattamente le specifiche ad una frequenza e le ecceda di molto alle altre. Questa affermazione intuitiva è confermata da un teorema che sarà discusso più oltre nel corso del paragrafo.

In tutte le discussioni che seguiranno tratteremo filtri FIR a fase nulla e con una risposta in frequenza della forma

$$H(e^{j\omega}) = \sum_{n=-M}^M h(n)e^{-j\omega n}$$

La durata della risposta all'impulso è $N=2M+1$, e per fase nulla imponiamo che sia

$$h(n) = h(-n)$$

Notiamo che un sistema causale può essere ottenuto semplicemente ritardando $h(n)$ di M campioni. A causa della simmetria di $h(n)$, possiamo scrivere $H(e^{j\omega})$ come

$$H(e^{j\omega}) = h(0) + \sum_{n=1}^M 2h(n) \cos(\omega n) \quad (5.59)$$

Dalla (5.59) osserviamo che $H(e^{j\omega})$ è puramente reale. Supponiamo di voler progettare un filtro passa-basso in accordo allo schema di tolleranze di fig. 5.42. Desideriamo cioè approssimare 1 nella banda $0 \leq |\omega| \leq \omega_p$ con errore massimo δ_1 e desideriamo approssimare zero nella banda $\omega_s \leq |\omega| \leq \pi$ con massimo errore δ_2 .

Non è naturalmente possibile specificare indipendentemente ciascuno dei parametri M , δ_1 , δ_2 , ω_p e ω_s . Sono stati tuttavia sviluppati degli algoritmi di progetto in cui alcuni di questi parametri sono fissati e i rimanenti vengono ottimizzati mediante una procedura iterativa. A questo proposito sono da considerare due diversi approcci. Herrmann e Schuessler [27, 28], e più tardi Hofstetter ed altri [29-31], hanno sviluppato procedure in cui sono fissati M , δ_1 e δ_2 , mentre ω_p e ω_s sono variabili. Parks e McClellan [32, 33] e Rabiner [34, 35] hanno sviluppato procedure in cui M , ω_p e ω_s sono fissati e δ_1 e δ_2 sono variabili. Poiché il lavoro di Herrmann e Schuessler è stato il primo proposto ed ha stimolato i successivi lavori sul progetto

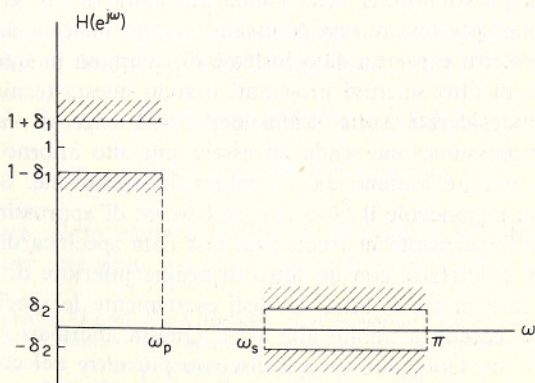


Fig. 5.42 Schema di tolleranze per l'approssimazione di un filtro passa-basso.

ad oscillazione uniforme, cominceremo la discussione partendo dal loro approccio.

Supponiamo di avere un'approssimazione ad oscillazione uniforme di un filtro passa-basso come quella rappresentata in fig. 5.43. Vedremo nel seguito che tali approssimazioni sono ottime nel senso di avere il più piccolo valore di δ_1 e δ_2 per valori fissati di ω_p e ω_s .

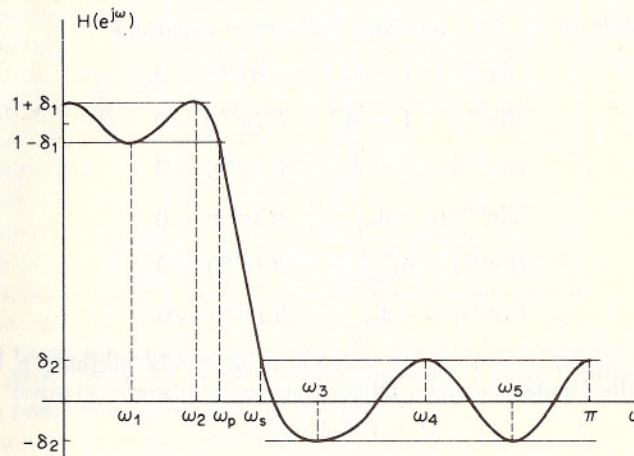


Fig. 5.43 Approssimazione ad oscillazione uniforme di un filtro passa-basso.

Un importante parametro dell'approssimazione ad oscillazione uniforme è il numero di massimi e minimi locali nel campo di frequenze $0 \leq \omega \leq \pi$. Per esaminare la dipendenza di questo parametro dalla lunghezza della risposta all'impulso, utilizziamo il fatto che $\cos \omega k$ può essere espresso come una somma di potenze di $\cos \omega$ per scrivere la (5.59) come

$$H(e^{j\omega}) = \sum_{k=0}^M a_k (\cos \omega)^k \quad (5.60)$$

dove gli a_k sono costanti legate ai valori della risposta all'impulso. La (5.60) mette in evidenza il fatto che $H(e^{j\omega})$ è un polinomio trigonometrico di ordine M . Come tale, può avere al massimo $M-1$ massimi e minimi locali nell'intervallo $0 < \omega < \pi$. Inoltre, se deriviamo la (5.60) rispetto ad ω otteniamo

$$H'(e^{j\omega}) = \frac{dH(e^{j\omega})}{d\omega} = -\sin \omega \left(\sum_{k=1}^M k a_k (\cos \omega)^{k-1} \right) \quad (5.61)$$

Dalla (5.61) notiamo che $H(e^{j\omega})$ avrà sempre un massimo o un minimo a $\omega = 0$ e $\omega = \pi$. Di conseguenza si avranno al massimo $M+1$ estremi locali di $H(e^{j\omega})$ nell'intervallo chiuso $0 \leq \omega \leq \pi$.

Utilizzando questo fatto Herrmann e Schuessler [27, 28] hanno mostrato come scrivere un sistema di equazioni che garantisce il comporta-

mento ad oscillazione uniforme riportato in fig. 5.43. I parametri M , δ_1 e δ_2 sono fissati ed ai parametri ω_p e ω_s è permesso di variare liberamente, essendo definiti dalle equazioni

$$H(e^{j\omega_p}) = 1 - \delta_1$$

$$H(e^{j\omega_s}) = \delta_2$$

Per l'esempio di fig. 5.43 possiamo scrivere le equazioni

$$H(e^{j0}) = 1 + \delta_1, \quad H(e^{j\pi}) = \delta_2$$

$$H(e^{j\omega_1}) = 1 - \delta_1, \quad H'(e^{j\omega_1}) = 0$$

$$H(e^{j\omega_2}) = 1 + \delta_1, \quad H'(e^{j\omega_2}) = 0$$

$$H(e^{j\omega_3}) = -\delta_2, \quad H'(e^{j\omega_3}) = 0$$

$$H(e^{j\omega_4}) = \delta_2, \quad H'(e^{j\omega_4}) = 0$$

$$H(e^{j\omega_5}) = -\delta_2, \quad H'(e^{j\omega_5}) = 0$$

In questo esempio vi sono tre estremi nella banda passante e quattro estremi nella banda attenuata. Di conseguenza abbiamo

$$M + 1 = 4 + 3 = 7$$

cioè $M = 6$ o $N = 13$. Vi sono $M + 1 = 7$ coefficienti incogniti nella (5.59) o (5.60), e cinque frequenze incognite $\omega_1, \omega_2, \dots, \omega_5$, in corrispondenza delle quali si hanno gli estremi, formando così un totale di 12 incognite che devono essere determinate come soluzione delle precedenti 12 equazioni. In generale possiamo avere N_p estremi nella banda passante e N_s estremi nella banda attenuata, dove

$$N_p + N_s = M + 1$$

e possiamo scrivere $2M$ equazioni che mettono in relazione gli $M + 1$ coefficienti del filtro e le $M - 1$ frequenze alle quali si hanno gli estremi (due estremi si hanno a $\omega = 0$ e $\omega = \pi$). Sfortunatamente queste equazioni sono non lineari e devono essere risolte utilizzando procedure iterative. A causa delle difficoltà numeriche che si incontrano risolvendo equazioni non lineari, questo approccio è risultato soddisfacente soltanto per bassi valori di M (dell'ordine di 30). Inoltre, per valori fissati di M , δ_1 e δ_2 , vi sono soltanto M differenti filtri ad oscillazione uniforme, in quanto si possono scegliere soltanto M differenti valori di N_p (o N_s). Ciò vuol dire che, per valori fissati di M , δ_1 e δ_2 , sono possibili soltanto M differenti scelte di ω_p . Di conseguenza, benché l'approssimazione ad oscillazione uniforme progettata con il metodo appena visto comporti la più stretta regione di transizione ($\omega_s - \omega_p$) per un dato valore di M , non si ha essenzialmente alcun miglioramento rispetto alla tecnica di progetto basata sul campionamento in frequenza, per quel che riguarda la scelta delle frequenze di taglio.

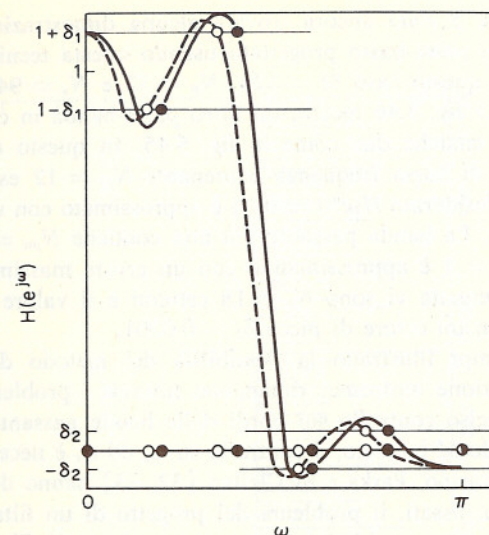


Fig. 5.44 Tipiche approssimazioni successive nella procedura di progetto di Hofstetter. La curva tratteggiata rappresenta l'approssimazione ottenuta interpolando i punti (circoletti) corrispondenti agli estremi della precedente approssimazione (curva a tratto intero) (da Hofstetter, ed altri [29]).

Una tecnica che affronta in maniera efficace le limitazioni connesse alla complessità dei calcoli nell'approccio di Herrmann e Schuessler è quella sviluppata da Hofstetter, Oppenheim e Siegel [29, 31], nella quale però permangono le restrizioni sulla scelta di ω_p e ω_s , essendo M , δ_1 e δ_2 fissate come in precedenza.

Invece di scrivere un sistema di equazioni non lineari, gli autori citati usano una tecnica iterativa per ottenere un polinomio trigonometrico che ha estremi del valore desiderato. La procedura inizia scegliendo N_p e N_s e quindi stimando le frequenze alle quali si verificano gli estremi. Si usano poi metodi standard di interpolazione di Lagrange [5] per calcolare un polinomio che assume i valori estremi prescritti ($1 \pm \delta_1$ nella banda passante e $\pm \delta_2$ nella banda attenuata) in corrispondenza delle frequenze stimate. Un esempio è mostrato dalla curva a tratto intero di fig. 5.44 per $N_p = N_s = 3$. I punti marcati rappresentano le stime iniziali delle frequenze di estremo. Gli estremi del polinomio risultante sono calcolati valutando il polinomio su di un insieme di frequenze finemente spaziate e ricercando su di questo i massimi ed i minimi locali. Se i massimi ed i minimi hanno i valori prescritti, la procedura è terminata; altrimenti viene calcolato un nuovo polinomio assumendo come nuove stime delle frequenze di estremo, gli estremi del precedente polinomio. La curva tratteggiata in fig. 5.44 mostra il polinomio risultante, dove i circoletti indicano le stime delle frequenze di estremo.

Il fatto che questa procedura converga all'approssimazione ad oscillazione uniforme desiderata è stato verificato per via numerica, ma di tale

convergenza non è stata ancora trovata alcuna dimostrazione [29]. Un esempio di filtro passa-basso progettato usando questa tecnica è mostrato in fig. 5.45. In questo caso $M = 125$, $N_p = 32$ e $N_s = 94$, $\delta_1 = 0.01$ e $\delta_2 = 0.00004$. La fig. 5.46 mostra un filtro passa-banda in cui vi sono tre regioni distinte anziché due come in fig. 5.45. In questo caso vi è una banda passante di bassa frequenza contenente $N_{pl} = 12$ estremi, e nella quale il valore desiderato $H(e^{j\omega}) = 0.25$ è approssimato con un errore massimo $\delta_{1l} = 0.01$. La banda passante più alta contiene $N_{pu} = 31$ estremi e il valore $H(e^{j\omega}) = 1$ è approssimativo con un errore massimo $\delta_{1u} = 0.02$. Nella banda attenuata vi sono $N_s = 18$ estremi e il valore $H(e^{j\omega}) = 0$ è approssimato con un errore di picco $\delta_2 = 0.0001$.

Questi esempi illustrano la flessibilità del metodo di approssimazione ad oscillazione uniforme; rimangono tuttavia i problemi della mancanza di un preciso controllo sui bordi delle bande passante e attenuata. Per avere, quando M è fissato, il controllo su ω_p ed ω_s è necessario consentire che δ_1 e δ_2 varino. Parks e McClellan [32, 33] hanno dimostrato che, con M , ω_p ed ω_s fissati, il problema del progetto di un filtro selettivo in frequenza diventa un problema di approssimazione di Chebyshev sopra insiemi disgiunti, che è un problema di notevole importanza nella teoria dell'approssimazione e per il quale pertanto sono già disponibili una quantità di utilissimi teoremi e procedimenti di calcolo [37]. Per formalizzare in questo caso il problema di approssimazione, definiamo la seguente funzione errore di approssimazione

$$E(\omega) = W(\omega)[H_d(e^{j\omega}) - H(e^{j\omega})] \quad (5.62)$$

da valutare sulle bande passante e attenuata del filtro desiderato; $W(\omega)$ è una funzione peso.

Supponiamo, per esempio, di voler ottenere un'approssimazione come in fig. 5.42, dove M , ω_p ed ω_s sono fissati. In tal caso,

$$H_d(e^{j\omega}) = \begin{cases} 1, & 0 \leq \omega \leq \omega_p \\ 0, & \omega_s \leq \omega \leq \pi \end{cases}$$

e

$$W(\omega) = \begin{cases} \frac{1}{K}, & 0 \leq \omega \leq \omega_p \\ 1, & \omega_s \leq \omega \leq \pi \end{cases}$$

La scelta di $W(\omega)$ specifica il rapporto fra le ampiezze degli errori di approssimazione della banda passante e di quella attenuata. Ciò implica che K sia uguale al rapporto desiderato δ_1/δ_2 . In questo caso il procedimento di progetto richiede un algoritmo per minimizzare il valore massimo di $|E(\omega)|$ negli intervalli $0 \leq \omega \leq \omega_p$ e $\omega_s \leq \omega \leq \pi$, il che equivale a minimizzare δ_2 .

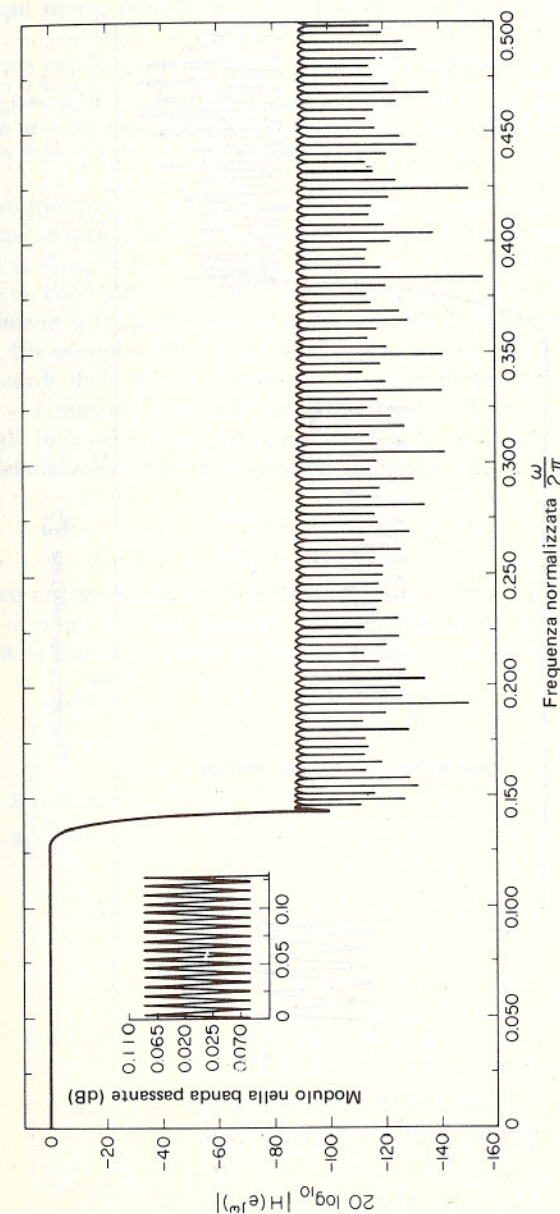


Fig. 5.45 Filtro passa-basso con $M = 125$, $N_p = 32$, $N_s = 94$, $\delta_1 = 0.01$ e $\delta_2 = 0.00004$ (da Hofstetter, ed altri [29]).

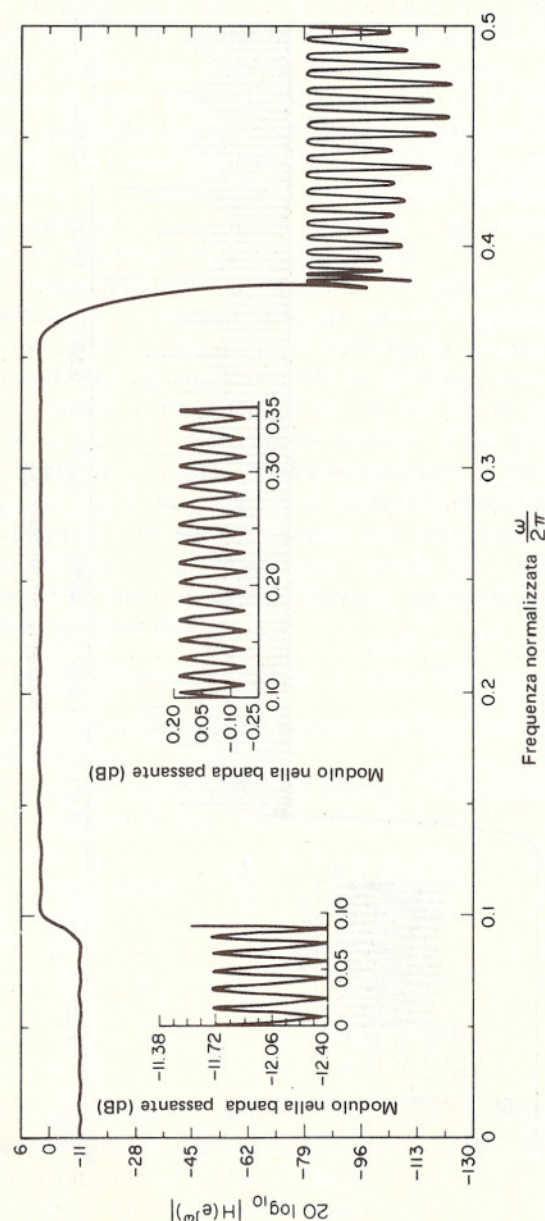


Fig. 5.46 Filtro passa-banda, con $M=60$, $N_{pi}=12$, $\delta_{li}=0.01$, $N_{pu}=31$, $\delta_{lu}=0.02$, $N_s=18$ e $\delta_s=0.0001$ (da Hofstetter, ed altri [29]).

Parks e McClellan [32] hanno riformulato un teorema di teoria dell'approssimazione in termini di teoria del progetto di filtri (secondo il metodo che qui stiamo discutendo), ottenendone il seguente teorema:

TEOREMA DELL'ALTERNANZA. Sia F un qualsiasi sottoinsieme chiuso dell'intervallo chiuso $0 \leq \omega \leq \pi$. Affinché $H(e^{j\omega})$ della (5.59) sia l'unica migliore approssimazione di $H_d(e^{j\omega})$ su F , è necessario e sufficiente che la funzione errore $E(\omega)$ abbia su F almeno $M+2$ « alternanze », per cui: $E(\omega_i) = -E(\omega_{i-1}) = \pm \|E\| = \max |E(\omega)|$ con $\omega_0 \leq \omega_1 \leq \omega_2 \leq \dots \leq \omega_{M+1}$ e ω_i contenuto in F .

È istruttivo interpretare questo teorema in termini del progetto di un filtro passa-basso. In questo caso il sottoinsieme chiuso F è costituito dagli intervalli $0 \leq \omega \leq \omega_p$ e $\omega_s \leq \omega \leq \pi$. Poiché $H_d(e^{j\omega})$ è costante a tratti, le frequenze ω_i corrispondenti a picchi nella funzione errore $E(\omega)$ corrispondono parimenti a frequenze per le quali $H(e^{j\omega})$ è al limite della tolleranza di errore. Un esempio tipico è schematizzato nella fig. 5.47.

Si ricordi dalla nostra precedente discussione che $H(e^{j\omega})$ può avere al più $M-1$ massimi e minimi locali nell'intervallo $0 < \omega < \pi$, e parimenti negli intervalli aperti combinati $0 < \omega < \omega_p$ e $\omega_s < \omega < \pi$. Inoltre, per definizione di banda passante e attenuata, $H(e^{j\omega})$ è vincolata dalle

$$H(e^{j\omega_p}) = 1 - \delta_1$$

$$H(e^{j\omega_s}) = +\delta_2$$

Ricordiamo anche che $H(e^{j\omega})$ avrà sempre un massimo o minimo locale in $\omega = 0$ e $\omega = \pi$. Pertanto possono esserci al massimo $M+3$ frequenze nelle quali la curva di errore raggiunge il suo massimo. Di conseguenza,

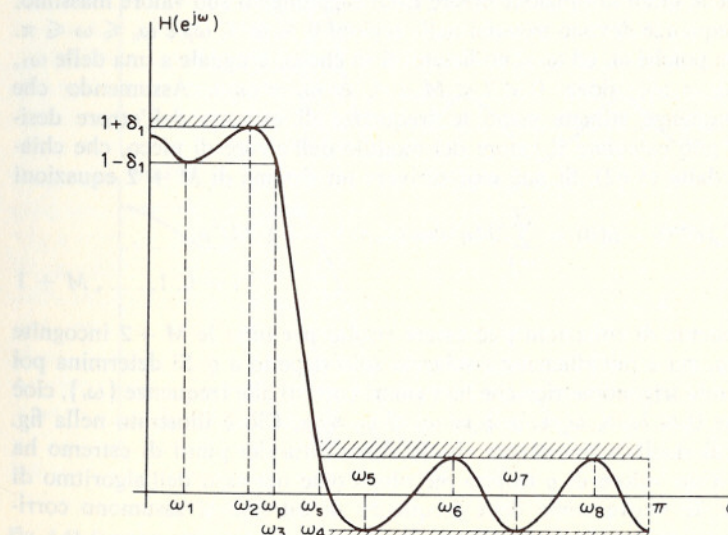


Fig. 5.47 Esempio tipico di approssimazione di filtro passa-basso che soddisfa il teorema dell'alternanza ($M=7$).

l'unica migliore approssimazione per la risposta passa-basso desiderata ha $M + 2$ oppure $M + 3$ alternanze della funzione errore. Nella fig. 5.48 sono mostrate quattro diverse possibili curve di risposta in frequenza per il valore $M = 7$. La fig. 5.48(a) mostra il caso in cui l'errore massimo si raggiunge sia per $\omega = 0$ che per $\omega = \pi$, e ci sono $M + 3$ alternanze. Le fig. 5.48(b) e (c) mostrano i casi in cui l'errore massimo si verifica rispettivamente solo in $\omega = \pi$ o solo in $\omega = 0$. In questi due casi si hanno solo $M + 2$ alternanze. La fig. 5.48(d) mostra il caso in cui si hanno solo $M + 2$ alternanze e l'errore massimo si raggiunge sia in $\omega = 0$ che in $\omega = \pi$. In accordo con il teorema dell'alternanza tutti questi filtri sono approssimazioni ottime⁸ per le prescritte frequenze di taglio ω_p (banda passante) e ω_s (banda attenuata). I filtri del tipo rappresentato nella fig. 5.48(a) sono stati chiamati da Parks e McClellan [32] filtri « con oscillazioni extra ». Questa terminologia è motivata dal fatto che tali filtri hanno un numero di alternanze di errore superiore al numero minimo ($M + 2$) richiesto per l'ottimalità. Se si includono i punti di estremo ω_p ed ω_s , i filtri progettati con le tecniche di Hermann-Schuessler e Hofstetter hanno $M + 3$ punti nei quali la risposta in frequenza raggiunge la tolleranza prescritta; pertanto tali filtri coincidono con i filtri con oscillazioni extra.

Oltre a stabilire delle chiare condizioni per l'ottimalità dei filtri FIR, Parks e McClellan [32] hanno anche presentato un procedimento iterativo per ottenere filtri ottimi. Questo procedimento è simile all'algoritmo di Hofstetter anche se in questo caso K , M , ω_p ed ω_s sono i parametri fissati e δ_2 è il parametro variabile. In accordo con il teorema dell'alternanza, il procedimento comincia con lo stimare $M + 2$ frequenze $\{\omega_i\}$, $i = 0, 1, \dots, M + 1$, nelle quali la funzione errore $E(\omega)$ raggiunge il suo valore massimo. Queste frequenze devono trovarsi nelle regioni $0 \leq \omega \leq \omega_p$ e $\omega_s \leq \omega \leq \pi$. Si noti che poiché ω_p ed ω_s sono fissate, si sa che ω_p è uguale a una delle ω_k , ovvero $\omega_p = \omega_l$, dove $0 < l < M + 1$ e $\omega_s = \omega_{l+1}$. Assumendo che queste frequenze stimate siano le frequenze di estremo dell'errore desiderato, si può calcolare il valore del modulo dell'errore di picco, che chiamiamo q , dalla (5.62). Si può cioè scrivere un sistema di $M + 2$ equazioni

$$W(\omega_i) \left[H_d(e^{j\omega_i}) - h(0) - \sum_{n=1}^M 2h(n) \cos(\omega_i n) \right] = -(-1)^i q, \\ i = 0, 1, \dots, M + 1$$

Questo sistema di equazioni può essere risolto per tutte le $M + 2$ incognite $\{h(n)\}$ e q , ma è più efficiente risolverlo solo rispetto a q . Si determina poi un polinomio trigonometrico che ha i valori corretti alle frequenze $\{\omega_i\}$, cioè $1 + Kq$ se $0 \leq \omega_i \leq \omega_p$ e $\pm q$ se $\omega_s \leq \omega_i \leq \pi$. Ciò è illustrato nella fig. 5.49 da cui risulta chiaramente che la stima fatta dei punti di estremo ha condotto a un valore di q troppo piccolo. Come nel caso dell'algoritmo di Hofstetter, le nuove stime delle frequenze di estremo si assumono corri-

⁸ Si intende *ottimo* nel senso della minimizzazione del massimo errore nella banda passante e nella banda oscura.

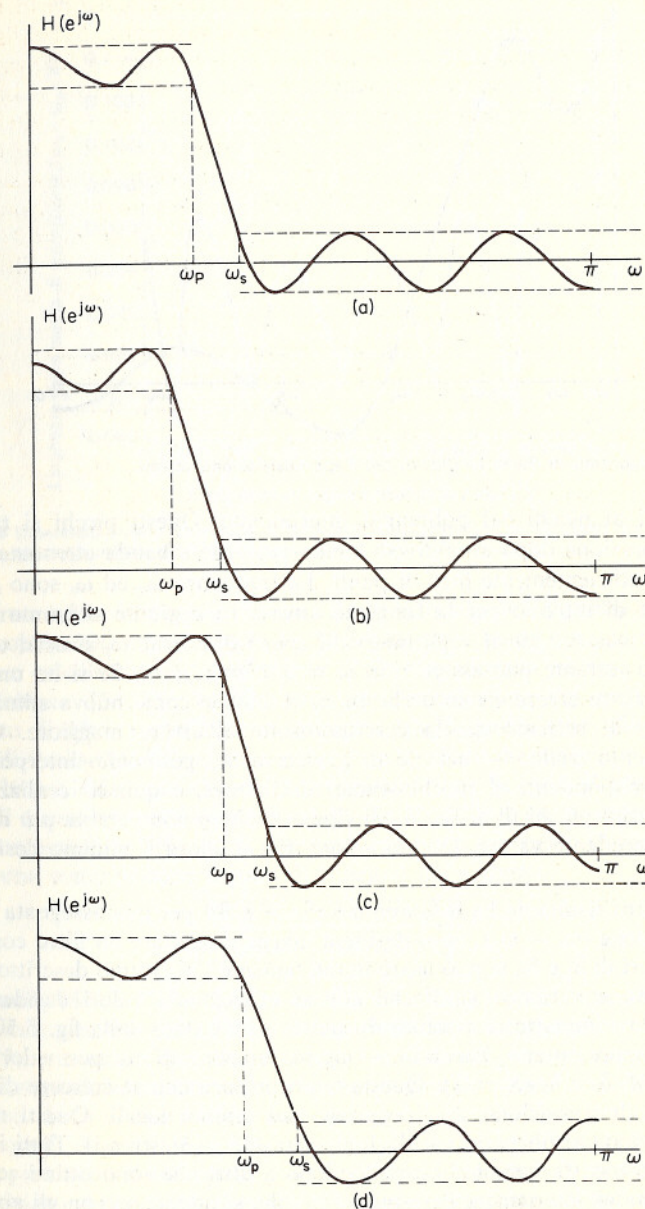


Fig. 5.48 Possibili approssimazioni ottime di filtri passa-basso per $M=7$: (a) $M+3$ alternanze (caso con oscillazioni extra); (b) $M+2$ alternanze (estremo in $\omega=\pi$); (c) $M+2$ alternanze (estremo in $\omega=0$); (d) $M+2$ alternanze (estremo sia in $\omega=0$ che in $\omega=\pi$).

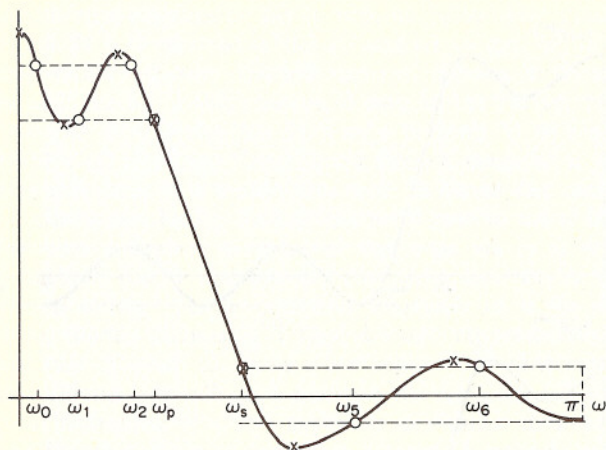


Fig. 5.49 Algoritmo di Parks-McClellan per l'approssimazione ottima.

spondenti ai picchi del polinomio interpolante. Questi picchi si trovano per esplorazione dopo aver diviso banda passante e banda attenuata in un insieme sufficientemente fitto di punti. Le frequenze ω_p ed ω_s sono ancora frequenze di estremo per la funzione errore. In aggiunta si hanno $M - 1$ minimi e massimi locali negli intervalli aperti $0 < \omega < \omega_p$ e $\omega_s < \omega < \pi$. L'estremo restante può essere o in $\omega = 0$, o in $\omega = \pi$. Se si ha un picco della funzione errore sia in 0 che in π , si assume come nuova stima della frequenza di estremo quella corrispondente all'errore maggiore. Questo procedimento ciclico — calcolo di q , ricerca del polinomio interpolante i punti corrispondenti ai picchi stimati dell'errore, e quindi localizzazione degli effettivi picchi di errore — si ripete finché q non cambia più rispetto al suo precedente valore. Questo valore di q è allora il minimo desiderato di δ_2 .

Il filtro risultante ha il δ_2 minimo ($\delta_1 = K \delta_2$) per una assegnata banda di transizione ($\omega_s - \omega_p$). Se si desidera invece progettare un filtro con assegnati valori di δ_1 e δ_2 , si può usare iterativamente l'algoritmo descritto sopra fissando ω_p e variando ω_s finché non si ottengono i valori desiderati di δ_1 e δ_2 . Un sommario di risultati di questo tipo è dato dalla fig. 5.50 dove è stato rappresentato $\Delta\omega = \omega_s - \omega_p$ in funzione di ω_p per valori assegnati di M , δ_2 e δ_1 ($K = 1$). Questa curva mostra che al crescere di ω_p la larghezza di transizione $\Delta\omega$ raggiunge dei minimi locali. Questi minimi corrispondono ai filtri con oscillazioni extra ($M + 3$ estremi). Tutti i punti che si trovano tra i minimi corrispondono a filtri che sono ottimi secondo il teorema dell'alternanza. Pertanto i filtri che si ottengono con gli approcci di Hermann-Schuessler e di Hofstetter sono casi particolari delle approssimazioni ottenute con l'algoritmo di Parks-McClellan.

Nell'algoritmo appena descritto tutti i valori della risposta all'impulso $h(n)$ vengono implicitamente variati ad ogni iterazione onde ottenere la

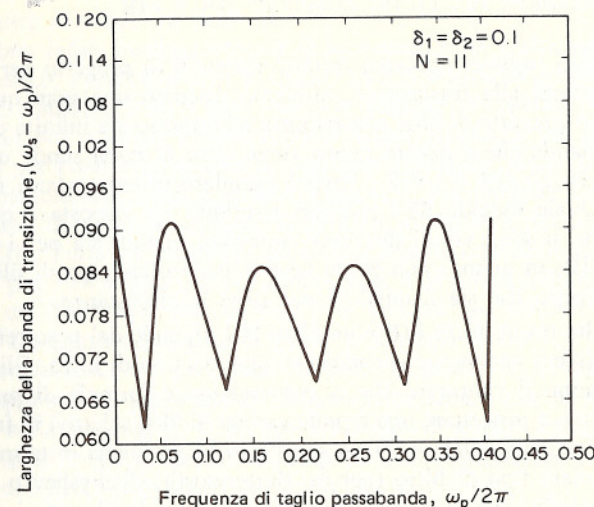


Fig. 5.50 Dipendenza della larghezza di transizione dalla frequenza di taglio per l'approssimazione ottima di un filtro passa-basso. $M=5$, $\delta_1=\delta_2=0.1$.

desiderata approssimazione ottima. Il passo finale dell'algoritmo è quello di calcolare $h(n)$ campionando in N o più punti la risposta in frequenza ottima e calcolando una trasformata di Fourier discreta inversa.

Rabiner [34, 35] ha discusso una formulazione del problema del progetto di filtri a oscillazione uniforme che è equivalente alla formulazione di Parks-McClellan, ma è strutturata in modo da consentire una soluzione basata sulle tecniche della programmazione lineare. Il punto essenziale di questo approccio sta nel notare che $H(e^{j\omega})$ può essere espressa come una combinazione lineare di funzioni coseno come nella (5.59), oppure come una combinazione lineare di funzioni della forma

$$\frac{\sin [N(\omega - (2\pi/N)k)/2]}{\sin [(\omega - (2\pi/N)k)/2]}$$

come nella (5.57). Nella (5.59) i coefficienti sono i valori $h(n)$ della risposta all'impulso, mentre nella (5.57) i coefficienti sono i campioni in frequenza $\tilde{H}(k)$. In entrambi i casi alcuni o tutti i coefficienti possono essere fatti variare sistematicamente in modo da raggiungere le prescritte tolleranze del progetto. Se si fanno variare solo alcuni dei parametri si ottengono filtri come quelli descritti nel paragrafo precedente. Se invece si fanno variare tutti i parametri, si ottengono le approssimazioni ottime. Il metodo di soluzione basato sulla programmazione lineare è più lento dal punto di vista dei calcoli, ma è più flessibile in quanto consente di operare con vincoli sia nel dominio temporale che in quello frequenziale [34, 35]. Maggiori dettagli sul progetto dei filtri FIR per risposte in frequenza comunque specificate possono trovarsi in [36].

5.7 UN CONFRONTO TRA FILTRI NUMERICI IIR E FIR

In questo capitolo abbiamo trattato i metodi di progetto per filtri numerici invarianti alla traslazione. Abbiamo discusso una gran quantità di metodi per il progetto di filtri con risposta all'impulso sia infinita che finita. Alcune domande che a questo punto viene fatto di porsi sono: quali filtri sono migliori, gli IIR o i FIR? Perché prendere in esame tanti metodi di progetto? Quale metodo dà i migliori risultati? La risposta a queste domande è che si sono voluti discutere numerosi metodi sia per i filtri IIR che per i FIR, in quanto non esiste nessun particolare tipo di filtro o metodo di progetto che sia il migliore per tutte le circostanze.

La scelta tra un filtro FIR e un filtro IIR dipende dal peso relativo che uno attribuisce a vantaggi e svantaggi di ciascun tipo di filtro. Gli IIR, per esempio, hanno il vantaggio che si possono usare formule di progetto in forma chiusa per progettare una grande varietà di filtri selettivi in frequenza. In altre parole, una volta che si sia specificato il problema in termini adatti a un particolare tipo di filtro (per es. Butterworth, Chebyshev o ellittico), si ha che i coefficienti (o i poli e gli zeri) del filtro desiderato si ottengono per sostituzione diretta in un insieme di equazioni di progetto. Questa semplificazione nel procedimento di progetto è vantaggiosa quando uno abbia da progettare solo pochissimi filtri, oppure quando le attrezzature di calcolo disponibili sono alquanto limitate.

Nel caso dei filtri FIR non esistono equazioni di progetto in forma chiusa. Il metodo delle finestre, pur potendosi applicare in maniera alquanto diretta e rapida, richiede spesso alcune iterazioni se si vogliono raggiungere determinate specifiche di progetto. La maggior parte degli altri metodi di progetto dei filtri FIR implicano procedimenti iterativi che richiedono mezzi di calcolo piuttosto potenti. Al contrario, è spesso possibile progettare filtri IIR selettivi in frequenza usando semplicemente un calcolatore tascabile e delle tavole con i parametri di progetto di filtri analogici. Il prezzo che si paga, tuttavia, per questa semplicità di progetto può essere misurato in termini di perdita di flessibilità rispetto al tipo di risposta di filtro ottenibile. I progetti di filtri IIR in forma chiusa sono essenzialmente limitati ai filtri passa-basso, passa-banda e passa-alto. Inoltre questi progetti non tengono generalmente conto della risposta di fase del filtro. Accade per esempio, che, benché si possa ottenere con una procedura di calcolo piuttosto semplice un filtro passa-basso di tipo ellittico con un'eccellente caratteristica di risposta in ampiezza, la sua risposta di fase risulti altamente non lineare (specialmente al bordo della banda).

Al contrario, i filtri FIR possono avere una fase esattamente lineare. Inoltre, il metodo delle finestre e la maggior parte dei metodi algoritmici consentono di approssimare risposte in frequenza alquanto arbitrarie con difficoltà di poco superiori a quelle che si incontrano nel progetto dei filtri passa-basso. A ciò si deve ancora aggiungere che, rispetto al caso dei filtri IIR, il problema del progetto dei filtri FIR è molto più sotto controllo grazie

all'esistenza di un teorema di ottimalità che risulta significativo per un ampio spettro di situazioni pratiche.

Esistono infine anche problemi di economia nella realizzazione di un filtro numerico. Preoccupazioni di questo tipo si misurano generalmente in termini di complessità di « hardware » e velocità di calcolo. Entrambi i fattori sono più o meno direttamente collegati all'ordine del filtro che è necessario per poter raggiungere determinate specifiche. Se si accantonano le considerazioni sulla fase, è generalmente vero che determinate specifiche sulla risposta di ampiezza si possono ottenere in maniera più efficiente con un filtro IIR. Tuttavia, in molti casi la fase lineare ottenibile mediante un filtro FIR può ben valere il costo addizionale, e in taluni altri casi [38] la scelta di un filtro FIR può perfino non comportare alcun sacrificio di efficienza. Un esame dettagliato di questi problemi è svolto in [39].

In conclusione, nel progettare un filtro numerico si devono considerare una quantità di pro e contro. È comunque chiaro che la scelta finale sarà fatta nella maggior parte dei casi sulla base di giudizi ingegneristici su problemi quali la formulazione delle specifiche, il metodo di realizzazione e i mezzi di calcolo disponibili per sviluppare il progetto.

SOMMARIO

In questo capitolo abbiamo preso in considerazione una certa quantità di tecniche di progetto di filtri numerici con risposta all'impulso di durata sia infinita che finita. Abbiamo posto l'accento sulle specifiche delle caratteristiche desiderate dei filtri nel dominio della frequenza, poiché questo è il caso più comune nella pratica. Il nostro obiettivo era di dare una visione complessiva del vasto campo di possibilità disponibili per il progetto di filtri numerici, fornendo nello stesso tempo una descrizione sufficientemente dettagliata di alcune delle tecniche, in modo che esse possano essere applicate direttamente, senza una ulteriore consultazione della letteratura specializzata. Perciò, abbiamo dedicato uno spazio considerevole ai metodi ormai acquisiti dell'invarianza all'impulso e della trasformazione bilineare ed abbiamo discusso in maniera molto meno completa alcuni dei metodi algoritmici di progetto di filtri IIR, poiché tali metodi sono molto meno usati nella pratica. Analogamente, abbiamo dedicato uno spazio considerevole al metodo delle finestre e a quello del campionamento in frequenza per il progetto di filtri FIR, mentre siamo stati più sintetici sui metodi algoritmici di progetto.

Il capitolo si chiude con alcune considerazioni sulla scelta tra le due classi di filtri numerici. Il punto centrale di questa discussione è che la scelta non sempre è ovvia e può dipendere da un gran numero di fattori difficili da valutare. Comunque, dovrebbe essere chiaro, da questo capitolo e dal precedente, che i filtri numerici sono caratterizzati da una grande flessibilità di progetto e di realizzazione; questo fatto rende possibile adottare schemi di elaborazione dei segnali piuttosto sofisticati, che in molti casi sarebbero difficili, se non impossibili, da realizzare con mezzi analogici.

BIBLIOGRAFIA

1. B. Gold and C. M. Rader, *Digital Processing of Signals*, McGraw-Hill Book Company, New York, 1969.
2. C. M. Rader and B. Gold, "Digital Filter Design Techniques in the Frequency Domain," *Proc. IEEE*, Vol. 55, Feb. 1967, pp. 149-171.
3. J. F. Kaiser, "Design Methods for Sampled Data Filters," *Proc. 1st Allerton Conf. Circuit System Theory*, Nov. 1963, pp. 221-236.
4. J. F. Kaiser, "Digital Filters," Chapter 7 in *System Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, John Wiley & Sons, Inc., New York, 1966.
5. F. B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill Book Company, New York, 1956.
6. A. J. Gibbs, "On the Frequency-Domain Responses of Causal Digital Filters," Ph.D. Thesis, University of Wisconsin, 1969.
7. E. A. Guillemin, *Synthesis of Passive Networks*, John Wiley & Sons, Inc., New York, 1957.
8. J. E. Storer, *Passive Network Synthesis*, McGraw-Hill Book Company, New York, 1957.
9. R. M. Golden and J. F. Kaiser, "Design of Wideband Sampled-Data Filters," *Bell System Tech. J.*, Vol. 43, No. 4, Pt. 2, July 1964, pp. 1533-1545.
10. A. Papoulis, "On the Approximation Problem in Filter Design," *IRE Conv. Record*, Pt. 2, 1957, pp. 175-185.
11. A. G. Constantinides, "Frequency Transformations for Digital Filters," *Elec. Lett.*, Vol. 3, No. 11, Nov. 1967, pp. 487-489.
12. A. G. Constantinides, "Frequency Transformations for Digital Filters," *Elec. Lett.*, Vol. 4, No. 7, Apr. 1968, pp. 115-116.
13. A. G. Constantinides, "Spectral Transformations for Digital Filters," *Proc. IEE*, Vol. 117, No. 8, Aug. 1970, pp. 1585-1590.
14. K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, June 1970.
15. L. R. Rabiner and K. Steiglitz, "The Design of Wide-Band Recursive and Nonrecursive Digital Differentiators," *IEEE Trans. Audio Electroacoust.*, Vol. 18, No. 2, June 1970, pp. 204-209.
16. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," *Computer J.*, Vol. 6, No. 2, 1963, pp. 163-168.
17. A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum P Error Criterion," *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, Oct. 1972, pp. 257-263.
18. C. S. Burrus and T. W. Parks, "Time Domain Design of Recursive Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, No. 2, June 1970, pp. 137-141.
19. J. L. Shanks, "Recursion Filters for Digital Processing," *Geophys.*, Vol. 32, No. 1, Feb. 1967, pp. 33-51.
20. N. Levinson, "The Wiener rms Error Criterion in Filter Design and Prediction," *J. Math. Phys.*, Vol. 25, No. 4, 1947, pp. 261-278.
21. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, Dover Publications, Inc., New York, 1958.
22. J. F. Kaiser, "Nonrecursive Digital Filter Design using the I_0 -sinh Window Function," *Proc. 1974 IEEE International Symp. on Circuits and Systems*, San Francisco, April, 1974, pp. 20-23.
23. B. Gold and K. L. Jordan, Jr., "A Direct Search Procedure for Designing Finite Duration Impulse Response Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-17, No. 1, Mar. 1969, pp. 33-36.
24. L. R. Rabiner, B. Gold, and C. A. McGonegal, "An Approach to the Approximation Problem for Nonrecursive Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, No. 2, June 1970, pp. 83-106.
25. L. R. Rabiner and R. W. Schafer, "Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," *IEEE Trans. Audio Electroacoust.*, Vol. 19, No. 3, Sept. 1971, pp. 200-207.
26. L. R. Rabiner and R. W. Schafer, "Correction to Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," *IEEE Trans. Audio Electroacoust.*, Vol. 20, No. 1, Mar. 1972, pp. 104-105.
27. O. Herrmann, "On the Design of Nonrecursive Digital Filters with Linear Phase," *Elec. Lett.*, Vol. 6, No. 11, 1970, pp. 328-329.
28. O. Herrmann and H. W. Schuessler, "Design of Nonrecursive Digital Filters with Minimum Phase," *Elec. Lett.*, Vol. 6, No. 11, 1970, pp. 329-330.
29. E. Hofstetter, A. V. Oppenheim, and J. Siegel, "A New Technique for the Design of Nonrecursive Digital Filters," *Proc. Fifth Annual Princeton Conf. Inform. Sci. Systems*, 1971, pp. 64-72.
30. E. Hofstetter, A. V. Oppenheim, and J. Siegel, "On Optimum Nonrecursive Digital Filters," *Proc. 9th Allerton Conf. Circuit System Theory*, Oct. 1971.
31. J. Siegel, "Design of Nonrecursive Approximations to Digital Filters with Discontinuous Frequency Responses," Ph.D. Thesis, MIT, June 1972.
32. T. W. Parks and J. H. McClellan, "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," *IEEE Trans. Circuit Theory*, Vol. CT-19, Mar. 1972, pp. 189-194.
33. T. W. Parks and J. H. McClellan, "A Program for the Design of Linear Phase Finite Impulse Response Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 3, Aug. 1972, pp. 195-199.
34. L. R. Rabiner, "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques," *Bell System Tech. J.*, July-Aug. 1972, pp. 1177-1198.
35. L. R. Rabiner, "Linear Program Design of Finite Impulse Response (FIR) Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. 20, No. 4, Oct. 1972, pp. 280-288.
36. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice Hall, Inc., Englewood Cliffs, N.J., 1975.
37. E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill Book Company, New York, 1966.
38. R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," *Proc. IEEE*, Vol. 61, No. 6, June 1973, pp. 692-702.
39. L. R. Rabiner, J. F. Kaiser, O. Herrmann, and M. T. Dolan, "Some Comparisons between FIR and IIR Digital Filters," *Bell Syst. Tech. J.*, vol. 53, No. 2, Febr., 1974, pp. 305-331.

PROBLEMI

1. Spesso si usano filtri numerici per elaborare segnali analogici limitati in banda nel modo illustrato in fig. P5.1. Nel caso ideale, il convertitore analogico-numerico campiona il segnale analogico fornendo la sequenza $x(n) = x_a(nT)$, ed il convertitore numerico-analogico trasforma i campioni $y(n)$ in una forma d'onda limitata in banda

$$y_a(t) = \sum_{n=-\infty}^{\infty} y(n) \frac{\sin(\pi/T)(t - nT)}{(\pi/T)(t - nT)}$$

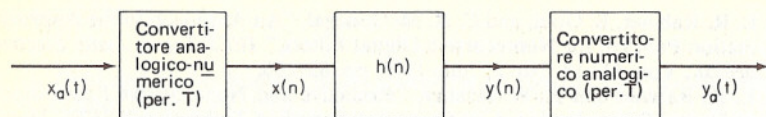


Fig. P5.1

Il sistema complessivo è equivalente ad un sistema analogico lineare tempo-invariante.

- Se il sistema $h(n)$ ha una frequenza di taglio di $\pi/8$ radianti e se $1/T = 10$ kHz, qual è la frequenza di taglio del filtro analogico equivalente?
 - Ripetere la parte (a) per $1/T = 20$ kHz.
2. L'approssimazione di un sistema analogico per mezzo di una equazione alle differenze all'indietro è stata discussa nel par. 5.1.2. Un procedimento alternativo potrebbe essere basato sulle differenze in avanti. Si assuma che $y(n) = y_a(nT)$ e $x(n) = x_a(nT)$ e si definisca la differenza prima in avanti come

$$\Delta^{(1)}[y(n)] = \frac{y(n+1) - y(n)}{T}$$

Le differenze di ordine superiore sono definite come

$$\Delta^{(k+1)}[y(n)] = \Delta^{(1)}[\Delta^{(k)}[y(n)]]$$

e

$$\Delta^{(0)}[y(n)] = y(n)$$

Si consideri l'approssimazione dell'equazione differenziale

$$\sum_{k=0}^N c_k \frac{d^k y_a(t)}{dt^k} = \sum_{k=0}^M d_k \frac{d^k x_a(t)}{dt^k}$$

con l'equazione alle differenze

$$\sum_{k=0}^N c_k \Delta^{(k)}[y(n)] = \sum_{k=0}^M d_k \Delta^{(k)}[x(n)]$$

- Se $H_a(s) = Y_a(s)/X_a(s)$ e $H(z) = Y(z)/X(z)$, determinare la funzione mappante $s = \psi(z)$ tale che

$$H(z) = H_a(\psi(z))$$

- Qual è il contorno nel piano z che è l'immagine dell'asse $j\Omega$ del piano s per la trasformazione ψ trovata nella parte (a)?
 - Sistemi stabili nel piano s sono mappati in sistemi stabili nel piano z ?
3. Siano $h_a(t)$, $s_a(t)$ e $H_a(s)$ la risposta all'impulso, la risposta al gradino e la funzione di trasferimento di un filtro a tempo continuo lineare tempo-invariante. Siano inoltre $h(n)$, $s(n)$ e $H(z)$ la risposta all'impulso, la risposta al gradino e la funzione di trasferimento di un filtro a tempo discreto lineare invariante alla traslazione.

- Se $h(n) = h_a(nT)$, è $s(n) = \sum_{k=-\infty}^n h_a(kT)$?

- Se $s(n) = s_a(nT)$, è $h(n) = h_a(nT)$?

4. Si consideri un sistema a tempo continuo con funzione di trasferimento

$$H_a(s) = \frac{s+a}{(s+a)^2 + b^2}$$

Determinare la funzione di trasferimento $H(z)$ (cioè la trasformata z della risposta all'impulso) di un sistema discreto progettato a partire da questo sistema sulla base di:

- Invarianza all'impulso, cioè in modo che sia $h(n) = h_a(nT)$

- Invarianza al gradino, cioè in modo che sia $s(n) = s_a(nT)$

dove

$$s(n) = \sum_{k=-\infty}^n h(k)$$

e

$$s_a(t) = \int_{-\infty}^t h_a(\tau) d\tau$$

5. Si supponga che $H_a(s)$ abbia un polo di ordine r in $s = s_0$, in modo che $H_a(s)$ si possa esprimere come

$$H_a(s) = \sum_{k=1}^r \frac{A_k}{(s-s_0)^k} + G_a(s)$$

dove $G_a(s)$ ha solo poli del primo ordine.

- Fornire una formula per determinare le costanti A_k da $H_a(s)$.
- Ottenere l'espressione della risposta all'impulso $h_a(t)$, in termini di s_0 e di $g_a(t)$, la trasformata di Laplace inversa di $G_a(s)$.
- Supponiamo di definire $h(n) = h_a(nT)$ come risposta all'impulso di un filtro numerico. Usando il risultato della parte (b), scrivere un'espressione della funzione di trasferimento $H(z)$.
- Discutere un procedimento diretto per ottenere $H(z)$ da $H_a(s)$.

6. Vogliamo progettare un filtro numerico passa-basso con una caratteristica di ampiezza nella banda passante che sia costante, entro 0.75 dB, per frequenze inferiori a $\omega = 0.2613\pi$, e con attenuazione nella banda oscura di almeno 20 dB per frequenze tra $\omega = 0.4018\pi$ e π .

Determinare la funzione di trasferimento $H(z)$ del filtro Butterworth di ordine minimo che soddisfa queste specifiche. Disegnare la realizzazione in cascata di questo filtro, includendo tutte le costanti necessarie. Usare la trasformazione bilineare.

Costanti Fondamentali

$\log^{-1}(0.75) = 5.6240$	$\log(\tan(0.4018\pi)) = 0.52840$
$\log^{-1}(0.075) = 1.1885$	$\log(525.2) = 2.72032$
$\tan(0.2613\pi) = 1.0736$	$\log(0.1885) = -0.72469$
$\log(\tan(0.2009\pi)) = -0.13616$	$\log(0.23721) = -0.6253$
$\log(\tan(0.1306\pi)) = -0.36131$	$\log(2.00000) = 0.30103$
$\log(\tan(0.2613\pi)) = 0.03081$	$\tan(0.1\pi) = 0.32492$

7. Molti filtri numerici IIR sono progettati mappando progetti analogici per mezzo della trasformazione bilineare. Le specifiche numeriche vengono trasformate in specifiche analogiche e quindi il corrispondente filtro analogico viene determinato o per mezzo di tavole di filtri, o con un programma al calcolatore o con calcoli a mano. Prima di usare una qualsiasi di tali tecniche, tuttavia, occorre determinare l'ordine N del filtro (N è uguale al numero dei poli). Proprio di ciò ci si occupa nel presente problema, che, nonostante la lunghezza del testo, è di rapida soluzione. Esaminate attentamente il problema da risolvere e l'informazione fornita. Troverete che vi è stata data molta più informazione del necessario. Un uso accurato delle tavole grafiche qui fornite consente inoltre di minimizzare i calcoli a mano (le fig. P5.7-3 ÷ P5.7-7 sono ristampate da [39]).

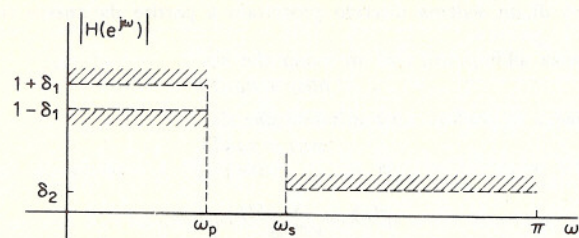


Fig. P5.7-1

Vogliamo progettare un filtro numerico che soddisfi le seguenti specifiche (v. fig. P5.7-1):

$$1 + \delta_1 \geq |H(e^{j\omega})| \geq 1 - \delta_1, \quad 0 \leq \omega \leq \omega_p$$

$$\delta_2 \geq |H(e^{j\omega})| \geq 0, \quad \omega_s \leq \omega \leq \pi$$

dove $\delta_1 = 0.01$, $\delta_2 = 0.01$, $\omega_p = 0.3\pi$ e $\omega_s = 0.4\pi$. Sfortunatamente, per ragioni storiche, i filtri analogici sono di norma specificati in termini di un diverso insieme di parametri e con un guadagno massimo unitario (fig. P5.7-2). Il vincolo sul guadagno massimo può essere superato moltiplicando le specifiche numeriche sull'ampiezza per un fattore $1/(1 + \delta_1)$. Perciò le specifiche numeriche ed analogiche sono legate da

$$\frac{1}{\sqrt{1 + \epsilon^2}} = \frac{1 - \delta_1}{1 + \delta_1}$$

e

$$\frac{1}{A} = \frac{\delta_2}{1 + \delta_1}$$

Ω_s , Ω_p , ω_s e ω_p sono legate dalla funzione distortrice della trasformazione bilineare.

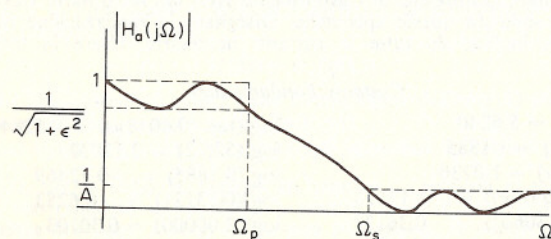


Fig. P5.7-2

È utile anche definire il parametro

$$\eta = \frac{\epsilon}{\sqrt{A^2 + 1}}$$

il quale è un parametro fondamentale per i filtri analogici usato nelle curve di progetto dei filtri stessi. Altre definizioni che potrebbero essere utili comprendono

$$k = \text{rapporto di transizione} = \frac{\Omega_p}{\Omega_s}$$

$$F_p = \text{frequenza di taglio per la banda passante} = \frac{\omega_p}{2\pi}$$

$$F_s = \text{frequenza di taglio per la banda oscura} = \frac{\omega_s}{2\pi}$$

$$\nu = F_s - F_p = \text{larghezza della banda di transizione}$$

$$N = \text{ordine del filtro}$$

Per i filtri ellittici,

$$N = \frac{K(k)K(\sqrt{1 - \eta^2})}{K(\eta)K(\sqrt{1 - k^2})},$$

dove $K(\cdot)$ è l'integrale ellittico completo di prima specie.

Per i filtri di Chebyshev,

$$N = \frac{\cosh^{-1}(1/\eta)}{\ln \beta},$$

dove

$$\beta = \frac{1 + \sqrt{1 - k^2}}{k}$$

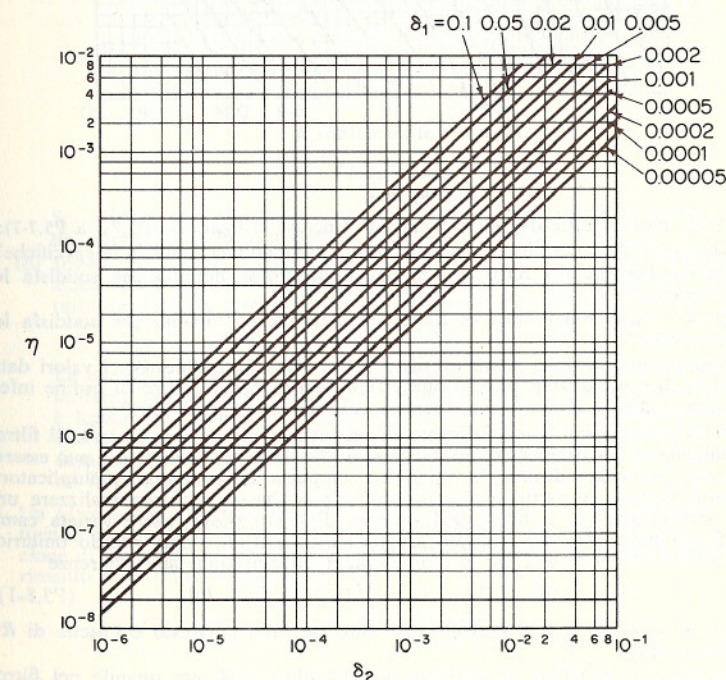
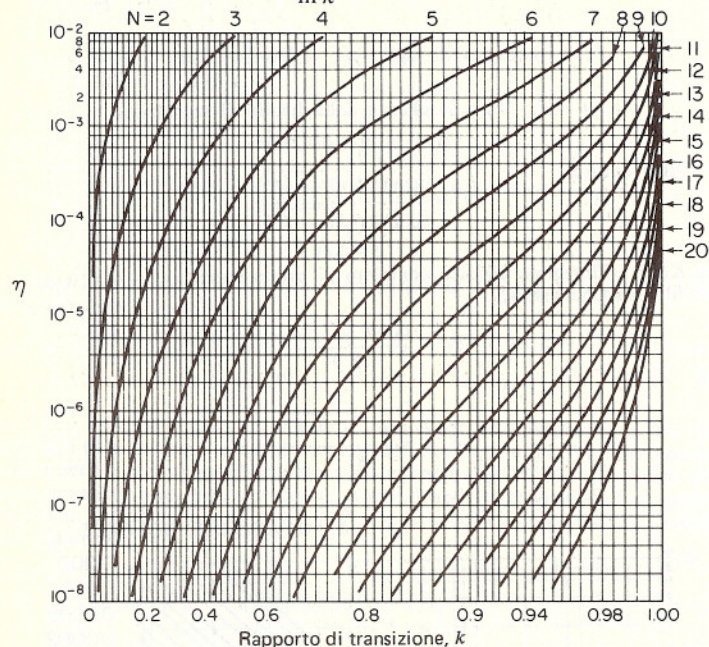


Fig. P5.7-3

Per i filtri di Butterworth,

$$N = \frac{\ln \eta}{\ln k}$$



Filtri ellittici

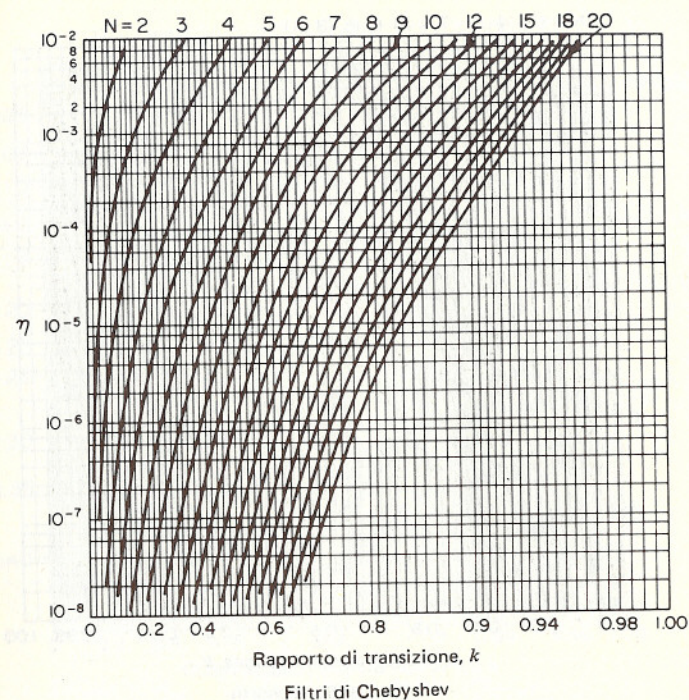
Fig. P5.7-4

- Sulla base di tutte queste informazioni (incluse le figure da P5.7-3 a P5.7-7):
- Qual è l'ordine del filtro ellittico di ordine minimo che soddisfa le specifiche?
 - Qual è l'ordine del filtro di Chebyshev di ordine minimo che soddisfa le specifiche?
 - Qual è l'ordine del filtro di Butterworth di ordine minimo che soddisfa le specifiche?
 - Supponiamo di voler usare un filtro del dodicesimo ordine. Per i valori dati di δ_1 , δ_2 e ω_p , quanto piccolo può essere reso ω_s con un filtro di ordine inferiore o uguale al dodicesimo per ciascuno dei tre tipi di filtri?
8. Sia $H(z)$ la funzione di trasferimento di un filtro numerico passa-basso. Il filtro corrisponde ad un sistema lineare, invariante alla traslazione, causale e può essere realizzato per mezzo di una rete numerica composta da sommatore, moltiplicatori e ritardi unitari. A partire da questo filtro passa-basso vogliamo realizzare un filtro passa-basso per il quale la frequenza di taglio possa essere variata cambiando un parametro. La strategia proposta è di sostituire ogni ritardo unitario nella rete di $H(z)$ con una rete R rappresentata dall'equazione alle differenze

$$y(n) = x(n-1) - \alpha[x(n) - y(n-1)] \quad (\text{P5.8-1})$$

dove α è reale, $|\alpha| < 1$, e $x(n)$ ed $y(n)$ rappresentano l'ingresso e l'uscita di R , rispettivamente.

- Sia $G(z)$ la funzione di trasferimento del filtro risultante quando nel filtro originario ogni ritardo unitario viene sostituito dalla rete corrispondente alla (P5.8-1). Mostrare che la risposta in frequenza associata a $G(z)$ è legata alla risposta in frequenza associata ad $H(z)$ da una trasformazione dell'asse frequenza; cioè se $G(e^{j\omega})$ e $H(e^{j\theta})$ indicano le due risposte in frequenza, ω può essere espressa come una funzione reale di θ . Disegnare ω in funzione di θ .



Filtri di Chebyshev

Fig. P5.7-5

Se la frequenza di taglio associata ad $H(z)$ è θ_p , determinare, come funzione del parametro α , la frequenza di taglio ω_p associata a $G(z)$.

- Invece di sostituire ciascun ritardo unitario con una rete descritta dall'equazione (P5.8-1), consideriamo il caso in cui ciascun ritardo unitario sia sostituito da una rete rappresentata dall'equazione alle differenze

$$y(n) = x(n-2) - \alpha[x(n-1) - y(n-1)] \quad (\text{P5.8-2})$$

La rete dell'equazione (P5.8-2) è quella della (P5.8-1) in cascata con un ritardo unitario. Anche in questo caso, si verificherà che la risposta associata a $G(z)$ è legata alla risposta in frequenza associata ad $H(z)$ da una trasformazione dell'asse frequenza. Determinare tale trasformazione e mostrare che in questo caso, se $H(z)$ corrisponde ad un filtro passa-basso, $G(z)$ non corrisponderà ad un filtro passa-basso.

- Un filtro analogico passa-alto può essere ottenuto da un filtro passa-basso sostituendo s con $1/s$ nella funzione di trasferimento; vale a dire, se $G_a(s)$ è la funzione di trasferimento per il filtro passa-basso, allora $H_a(s)$ è la funzione di trasferimento di un filtro passa-alto se vale

$$H_a(s) = G_a\left(\frac{1}{s}\right)$$

D'altra parte un filtro numerico può essere ottenuto attraverso la trasformazione di un filtro analogico per mezzo della trasformazione bilineare

$$s = \frac{z-1}{z+1}$$

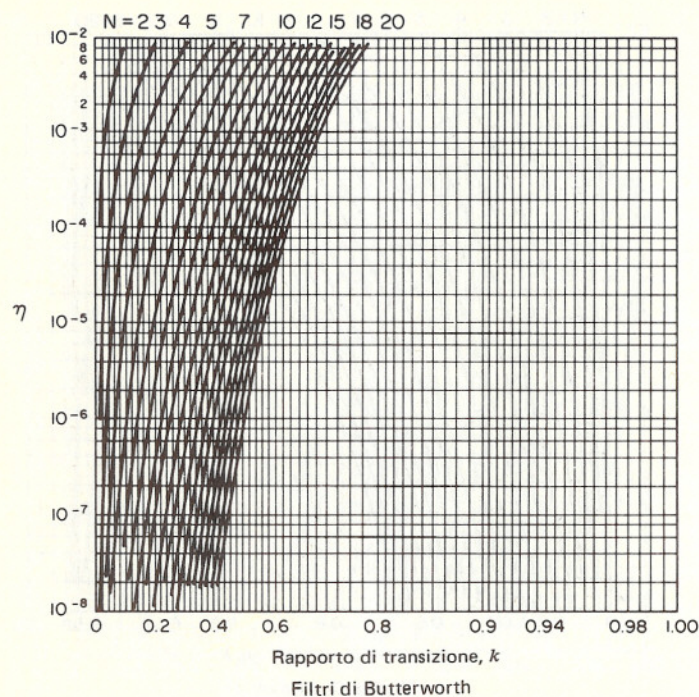


Fig. P5.7-6

Si è assunto, per comodità, $T = 2$ nella relazione (5.22). Questa trasformazione conserva l'andamento (per es. passa-basso) del modulo della funzione di trasferimento, anche se la scala delle frequenze risulta distorta. La rete di fig. P5.9 rappresenta un filtro passa-basso con frequenza di taglio $\omega_c = \pi/2$.

Le costanti A , B , C e D sono reali. Determinare la maniera in cui vanno modificati i coefficienti per ottenere un filtro passa-alto con frequenza di taglio $\omega_H = \pi/2$.

10. Si assuma che il filtro a tempo continuo sia del tipo passa-basso e che $H(z) = H_c((z+1)/(z-1))$. Allora la banda passante del filtro numerico sarà centrata in:
- $\omega = 0$ (passa-basso)
 - $\omega = \pi$ (passa-alto)
 - una frequenza diversa da 0 o π (passa-banda)
- (Scegliere la risposta corretta).
11. Sia $h(n)$ la risposta all'impulso di un filtro FIR tale che $h(n) = 0$ per $n < 0$, $n \geq N$. Si assuma $h(n)$ reale. È possibile garantire che il filtro abbia fase lineare imponendo certe condizioni di simmetria sulla sua risposta all'impulso $h(n)$.
- La risposta in frequenza di questo filtro può essere espressa nella forma

$$H(e^{j\omega}) = \hat{H}(e^{j\omega}) e^{j\theta(\omega)}$$

dove

$$\hat{H}(e^{j\omega}) \text{ è reale.}$$

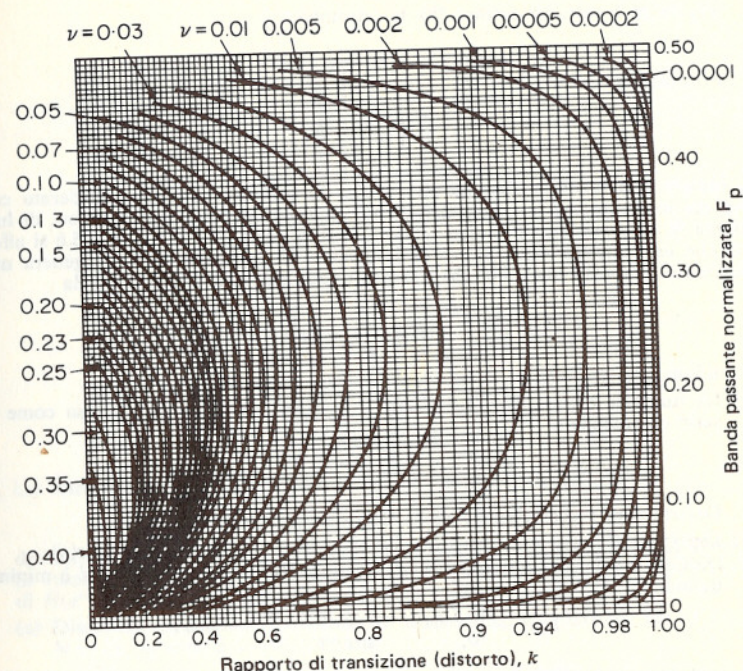


Fig. P5.7-7

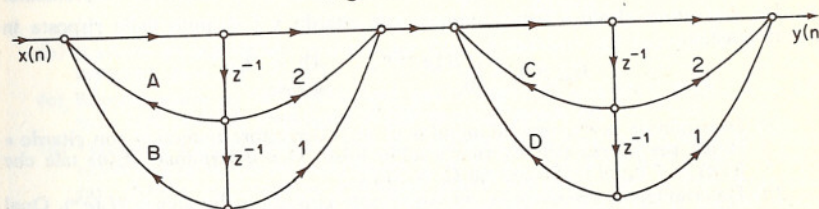


Fig. P5.9

- Trovare $\theta(\omega)$, per $0 \leq \omega \leq \pi$, quando $h(n)$ soddisfa la condizione

$$h(n) = h(N-1-n)$$
- Trovare $\theta(\omega)$, per $0 \leq \omega \leq \pi$, quando

$$h(n) = -h(N-1-n)$$

(Si faccia molta attenzione: può essere necessario trattare separatamente il caso di N dispari e quello di N pari).

- Si indichi con $H(k)$ la DFT di $h(n)$ su N punti.

- Se $h(n)$ soddisfa la relazione

$$h(n) = -h(N-1-n)$$

dimostrare che vale

$$H(0) = 0$$

(2) Se N è pari, dimostrare che la relazione

$$h(n) = h(N - 1 - n)$$

implica

$$H\left(\frac{N}{2}\right) = 0$$

12. Si indichi con $h_d(n)$ la risposta all'impulso di un sistema ideale desiderato con risposta in frequenza $H_d(e^{j\omega})$ e si indichi con $h(n)$ la risposta all'impulso, di lunghezza N , di un sistema FIR con risposta in frequenza $H(e^{j\omega})$. Nel par. 5.6 si affermò che una finestra rettangolare di lunghezza N applicata ad $h_d(n)$ genera una risposta all'impulso $h(n)$ tale che l'errore quadratico medio, espresso da

$$\varepsilon^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_d(e^{j\omega}) - H(e^{j\omega})|^2 d\omega$$

è minimizzato.

- (a) La funzione errore $E(e^{j\omega}) = H_d(e^{j\omega}) - H(e^{j\omega})$ può essere espressa come la serie di potenze

$$E(e^{j\omega}) = \sum_{n=-\infty}^{\infty} e(n)e^{-j\omega n}$$

Determinare i coefficienti $e(n)$ in funzione di $h_d(n)$ e di $h(n)$.

- (b) Esprimere l'errore quadratico medio ε^2 in funzione dei coefficienti $e(n)$.
(c) Dimostrare che per una risposta all'impulso $h(n)$ di lunghezza N , ε^2 è minimo quando

$$h(n) = \begin{cases} h_d(n), & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

Cioè la finestra rettangolare dà la migliore approssimazione secondo i minimi quadrati di una risposta in frequenza desiderata, per un valore di N prefissato.

13. Un derivatore ideale a banda limitata con ritardo τ è definito dalla risposta in frequenza

$$H_d(j\Omega) = \begin{cases} (j\Omega)e^{-j\Omega\tau}, & |\Omega| \leq \Omega_c \\ 0, & \text{altrove} \end{cases}$$

- (a) Determinare la risposta all'impulso di un « derivatore numerico con ritardo » ideale per mezzo dell'invarianza all'impulso; cioè determinare $h_d(n)$ tale che $h_d(n) = T h_d(nT)$. (Si assuma $\Omega_c = \pi/T$).
(b) Determinare e disegnare la corrispondente risposta in frequenza $H_d(e^{j\omega})$. Qual è il ritardo del sistema in numero di campioni?
(c) Una maniera per ottenere un'approssimazione causale al derivatore numerico è quella di usare il metodo delle finestre. Si supponga che la risposta all'impulso $h(n)$ di tale approssimazione sia diversa da zero solo nell'intervallo $0 \leq n \leq N-1$. Come bisogna scegliere τ nel caso che N sia pari (1) o che N sia dispari (2)? Qual è il ritardo espresso in numero di campioni nei due casi? Si disegni una tipica risposta all'impulso per ogni caso.
(d) Si ponga $N = 2$ e $\tau = T/2$ e si scelga

$$h(n) = \begin{cases} h_d(n), & n = 0, 1 \\ 0, & \text{altrove} \end{cases}$$

- (1) Esprimere l'uscita di questo sistema in funzione dell'ingresso.
(2) Qual è la risposta in frequenza in questo caso?
(3) Ottenere un'espressione per l'errore relativo

$$E_r(\omega) = \frac{H_d(e^{j\omega}) - H(e^{j\omega})}{H_d(e^{j\omega})}$$

per questo caso. Fare un grafico di $E_r(\omega)$ per $0 \leq \omega \leq \pi$.

- (e) Ripetere il punto (d) ponendo $N = 3$ e $\tau = T$.

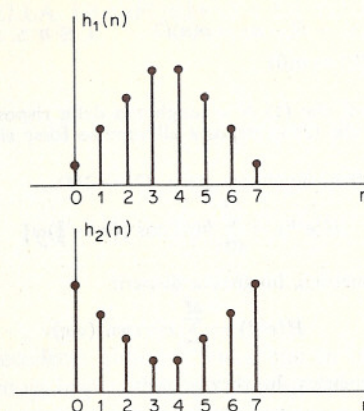


Fig. P5.15

14. Un filtro con risposta all'impulso finita ha risposta in frequenza

$$H(e^{j\omega}) = |H(e^{j\omega})|e^{-j\omega n_0}$$

dove n_0 non è necessariamente intero. Sia N la lunghezza della risposta all'impulso. Si ricordi che la risposta all'impulso è completamente definita da N campioni di $H(e^{j\omega})$ presi in $\omega = 2\pi k/N$, $k = 0, 1, \dots, N-1$.

- (a) Disegnare approssimativamente il diagramma di $|H(e^{j\omega})|$ per il caso di $N = 15$, $n_0 = 0$, e

$$|\tilde{H}(k)| = |H(e^{j(2\pi/15)k})| = \begin{cases} 1, & k = 0 \\ \frac{1}{2}, & k = 1, 14 \\ 0, & \text{altrove} \end{cases}$$

- (b) Scrivere un'espressione generale per $h(n)$ in funzione dei termini $\tilde{H}(k)$. (Non assumere $n_0 = 0$).
(c) Rappresentare in un grafico $h(n)$ per i casi: (1) $n_0 = (N-1)/2 = 7$, e (2) $n_0 = N/2 = 15/2$, scegliendo $|\tilde{H}(k)|$ come al punto (a).
(d) Disegnare un diagramma a blocchi completo (cioè una rete numerica) di una realizzazione di questo sistema con $N = 15$, $n_0 = 15/2$ e $|\tilde{H}(k)|$ come al punto (a). Questa realizzazione dovrebbe essere ricorsiva, vale a dire dovrebbe essere basata sul campionamento in frequenza. Si confronti il numero delle addizioni e moltiplicazioni richieste con quello necessario per una realizzazione in forma diretta.
15. Si considerino due sequenze di durata finita, $h_1(n)$ e $h_2(n)$, di lunghezza 8, rappresentate in fig. P5.15: esse sono legate da una rotazione circolare.
- (a) I moduli delle due DFT su otto punti sono uguali. (È vero o è falso?)
(b) Si voglia realizzare un filtro passa-basso non ricorsivo e si debba utilizzare come risposta all'impulso $h_1(n)$ o $h_2(n)$. Quale delle seguenti affermazioni è giusta?
- (1) $h_1(n)$ è un filtro passa-basso migliore di $h_2(n)$.
 - (2) $h_2(n)$ è un filtro passa-basso migliore di $h_1(n)$.
 - (3) Sono entrambi ugualmente buoni (o cattivi) filtri passa-basso.
16. Nella sua forma originaria, l'algoritmo di Parks-McClellan determinava la sequenza $a(n)$ tale che

$$\min_{\{a(n)\}} \left\{ \max_{\omega} \left[W(\omega) \left| H_d(e^{j\omega}) - \sum_{n=0}^M a(n) \cos \omega n \right| \right] \right\} \quad (\text{P5.16-1})$$

dove

$$\begin{aligned} h(n) &= h(-n) = a(n)/2 & 1 \leq n \leq M \\ h(0) &= a(0) \end{aligned}$$

Esso richiedeva quindi che (1) N = lunghezza della risposta all'impulso = $2M + 1$ fosse dispari, e che (2) la risposta all'impulso fosse simmetrica. Ci sono altri tre casi interessanti:

A: Risposta simmetrica, lunghezza pari: ($N = 2M$)

$$H(e^{j\omega}) = \sum_{n=1}^M b(n) \cos[(n - \frac{1}{2})\omega]$$

B: Risposta antisimmetrica, lunghezza dispari:

$$H(e^{j\omega}) = \sum_{n=1}^M c(n) \sin(\omega n)$$

C: Risposta antisimmetrica, lunghezza pari:

$$H(e^{j\omega}) = \sum_{n=1}^M d(n) \sin(\omega(n - \frac{1}{2}))$$

Dato un algoritmo che calcola $a(n)$ secondo l'espressione (P5.16-1), mostrare come esso può essere usato per progettare filtri che corrispondono agli altri tre casi.

- Trovare una funzione $W(e^{j\omega})$ che può essere usata per risolvere il caso A e mostrare come si può determinare $b(n)$ a partire da $a(n)$.
- Trovare una funzione $W(e^{j\omega})$ che può essere usata per risolvere il caso B e mostrare come si può determinare $c(n)$ a partire da $a(n)$.
- Trovare una funzione $W(e^{j\omega})$ che può essere usata per risolvere il caso C e mostrare come si può determinare $d(n)$ a partire da $a(n)$.

6. CALCOLO DELLA TRASFORMATATA DI FOURIER DISCRETA

6.0 INTRODUZIONE

Nei capitoli precedenti abbiamo visto che la trasformata di Fourier discreta ricopre un ruolo importante nell'analisi, nel progetto e nella realizzazione di algoritmi e sistemi per l'elaborazione numerica dei segnali. Ulteriore conferma di questa affermazione si avrà nei capitoli successivi. Uno dei motivi per cui l'analisi di Fourier è così importante e di vasto impiego nell'elaborazione numerica dei segnali è l'esistenza di algoritmi efficienti per il calcolo della trasformata di Fourier discreta [1].

Ricordiamo dal cap. 3 che la trasformata di Fourier discreta (DFT) è definita come

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad k = 0, 1, \dots, N-1 \quad (6.1)$$

con $W_N = e^{-j(2\pi/N)}$. La trasformata di Fourier discreta inversa è

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}, \quad n = 0, 1, \dots, N-1 \quad (6.2)$$

Nelle (6.1) e (6.2) sia $x(n)$ che $X(k)$ possono essere complesse. Le due espressioni differiscono solo per il segno dell'esponente di W_N e per un fattore scala pari a $1/N$. Perciò la discussione delle procedure di calcolo per la (6.1) vale anche, con semplici modifiche, per la (6.2).

Per dimostrare l'importanza di disporre di schemi di calcolo efficienti, è istruttivo esaminare il caso del calcolo diretto della DFT, cioè delle espressioni (6.1) e (6.2). Poiché $x(n)$ è in generale complessa si può scrivere

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} \{(\operatorname{Re}[x(n)] \operatorname{Re}[W_N^{kn}] - \operatorname{Im}[x(n)] \operatorname{Im}[W_N^{kn}]) \\ &\quad + j(\operatorname{Re}[x(n)] \operatorname{Im}[W_N^{kn}] + \operatorname{Im}[x(n)] \operatorname{Re}[W_N^{kn}])\}, \\ &\quad k = 0, 1, \dots, N-1 \quad (6.3) \end{aligned}$$

Dall'espressione precedente risulta chiaro che il calcolo diretto di $X(k)$ richiede $4N$ moltiplicazioni reali e $(4N-2)$ addizioni reali per ogni valore di k ¹. Poiché occorre valutare $X(k)$ per N diversi valori di k , il calcolo

¹ In tutta questa discussione il numero di operazioni è solo approssimato. La moltiplicazione per W_N , ad es., non richiede in realtà una moltiplicazione. Ciò nonostante, la dipendenza generale della complessità di calcolo dal valore di N che si ottiene includendo queste moltiplicazioni è abbastanza precisa per consentire confronti tra diverse classi di algoritmi.

diretto della trasformata di Fourier discreta di una sequenza $x(n)$ richiede $4N^2$ moltiplicazioni reali e $N(4N - 2)$ addizioni reali ovvero, in altri termini, N^2 moltiplicazioni complesse e $N(N - 1)$ addizioni complesse. Oltre alle moltiplicazioni e alle addizioni contenute nella (6.3), l'esecuzione del calcolo della DFT su un calcolatore numerico d'impiego generale o con un dispositivo ad essa dedicato implica naturalmente la necessità di memorizzare e di leggere i valori della sequenza d'ingresso $x(n)$ e dei coefficienti W_N^{kn} . Poiché negli algoritmi di calcolo numerico la quantità di operazioni di lettura e scrittura risulta, in generale, proporzionale al numero di operazioni aritmetiche, viene comunemente accettato come misura significativa della complessità (ovvero del tempo richiesto per eseguire un algoritmo di calcolo) il numero di moltiplicazioni e di addizioni che sono necessarie. Quindi, per il calcolo diretto della trasformata di Fourier discreta, l'efficienza del metodo si può valutare sulla base del fatto che sono necessarie $4N^2$ moltiplicazioni reali e $N(4N - 2)$ addizioni reali. Poiché la quantità (e quindi il tempo) dei calcoli è approssimativamente proporzionale a N^2 , è evidente che il numero di operazioni aritmetiche richiesto per calcolare la DFT con il metodo diretto diventa enorme per valori grandi di N . Per questo motivo risultano di grande interesse procedure di calcolo che riducano il numero di moltiplicazioni e addizioni.

La maggior parte delle tecniche usate per migliorare l'efficienza del calcolo della DFT sfruttano almeno una delle seguenti proprietà particolari delle quantità W_N^{kn} :

1. $W_N^{k(N-n)} = (W_N^{kn})^*$.
2. $W_N^{kn} = W_N^{k(n+N)} = W_N^{(k+N)n}$

Per esempio, usando la prima proprietà, cioè la simmetria delle funzioni seno e coseno, si possono nella (6.3) raggruppare dei termini come

$$\begin{aligned} \operatorname{Re}[x(n)] \operatorname{Re}[W_N^{kn}] + \operatorname{Re}[x(N-n)] \operatorname{Re}[W_N^{k(N-n)}] \\ = (\operatorname{Re}[x(n)] + \operatorname{Re}[x(N-n)]) \operatorname{Re}[W_N^{kn}] \end{aligned}$$

e

$$\begin{aligned} -\operatorname{Im}[x(n)] \operatorname{Im}[W_N^{kn}] - \operatorname{Im}[x(N-n)] \operatorname{Im}[W_N^{k(N-n)}] \\ = -(\operatorname{Im}[x(n)] - \operatorname{Im}[x(N-n)]) \operatorname{Im}[W_N^{kn}] \end{aligned}$$

Analoghi raggruppamenti si possono fare per gli altri termini della (6.3). In questo modo si riesce a ridurre il numero di moltiplicazioni circa di un fattore 2. Si può anche sfruttare il fatto che per certi valori del prodotto kn le funzioni seno e coseno valgono 1 o 0, e non è quindi necessario eseguire le moltiplicazioni corrispondenti. Tuttavia, dopo riduzioni di questo genere, resta sempre da eseguire una quantità di calcoli che è all'incirca proporzionale a N^2 . Per fortuna, è possibile usare la seconda proprietà, cioè la periodicità della sequenza complessa W_N^{kn} , per ottenere una riduzione dei calcoli notevolmente maggiore.

Algoritmi di calcolo che sfruttano sia la simmetria che la periodicità della sequenza W_N^{kn} erano noti già da molto tempo prima dell'avvento dei calcolatori numerici veloci. A quei tempi, qualunque accorgimento che riducesse i calcoli, fatti a mano, anche solo di un fattore 2 era visto con interesse. Runge [2] e più tardi Danielson e Lanczos [3] hanno descritto algoritmi la cui complessità era all'incirca proporzionale a $N \log N$ invece che a N^2 . Questa differenza non era però di grande importanza essendo piccoli i valori di N che permettono di fare i conti a mano.² La possibilità di ridurre notevolmente i tempi di calcolo non divenne chiara se non verso il 1965, quando Cooley e Tukey [1] pubblicarono un algoritmo per il calcolo della trasformata di Fourier discreta che vale quando N è un numero composto, cioè il prodotto di due o più interi. La pubblicazione di questo lavoro provocò un fiorire di applicazioni della trasformata di Fourier discreta all'elaborazione dei segnali e diede luogo alla scoperta di numerosi algoritmi di calcolo che divennero noti come *algoritmi per la trasformata di Fourier veloce*, o semplicemente FFT. Sinteticamente, l'intero insieme di questi algoritmi è spesso indicato come « la FFT » [5].

Il principio fondamentale su cui si basano tutti questi algoritmi è la scomposizione del calcolo della trasformata di Fourier discreta di una sequenza lunga N in trasformate di Fourier discrete di dimensioni via via più piccole. Il modo in cui questo principio è applicato dà luogo a una varietà di algoritmi diversi, tutti caratterizzati da miglioramenti circa della stessa entità nella velocità di calcolo. In questo capitolo ci occuperemo di due classi fondamentali di algoritmi di FFT. La prima, detta a *decimazione nel tempo*, prende il nome dal fatto che nello scomporre il calcolo in trasformate di dimensioni più piccole, la sequenza $x(n)$ (l'indice n è spesso associato al tempo) viene suddivisa in sequenze sempre più corte. Nella seconda classe generale di algoritmi, la sequenza che viene scomposta in sottosequenze sempre più corte è quella dei coefficienti della trasformata di Fourier discreta, da cui il nome di *decimazione in frequenza*.

In questo capitolo prenderemo in esame un certo numero di algoritmi per calcolare la trasformata di Fourier discreta. Questi algoritmi, pur avendo diversa efficienza, sono comunque tutti più efficienti che non il calcolo diretto dell'espressione (6.3). Cominceremo discutendo il metodo di Goertzel [6, 7], che richiede un numero di operazioni proporzionale a N^2 ma con una costante di proporzionalità più piccola rispetto al metodo diretto. La nostra attenzione sarà poi dedicata in massima parte a discutere gli algoritmi di FFT, cioè algoritmi per i quali l'ammontare dei calcoli è circa proporzionale a $N \log N$. Nella presentazione di questi metodi non cercheremo di essere esaustivi, ma illustreremo i principi generali comuni a tutti gli algoritmi di questo tipo, considerando in dettaglio solo alcune delle tecniche più comunemente usate.

² Un interessante articolo di Cooley, Lewis e Welch [4] presenta la storia degli sforzi volti a migliorare l'efficienza del calcolo della DFT.

6.1 L'ALGORITMO DI GOERTZEL

L'algoritmo di Goertzel [6] è un procedimento di calcolo più efficiente del metodo diretto ed è un esempio di come si possa sfruttare la periodicità della sequenza W_N^{kn} per ridurre i calcoli. Più precisamente, vedremo che la trasformata di Fourier discreta può essere considerata come la risposta di un filtro numerico la cui struttura può essere progettata in modo da ridurre il numero delle operazioni aritmetiche.

Per ricavare l'algoritmo di Goertzel, cominciamo col notare che è

$$W_N^{-kN} = e^{j(2\pi/N)Nk} = e^{j2\pi k} = 1 \quad (6.4)$$

Questa è, ovviamente, una conseguenza immediata della periodicità di W_N^{-kn} . In base alla (6.4) possiamo moltiplicare il secondo membro della relazione (6.1) per W_N^{-kN} senza alterare l'uguaglianza. Quindi è

$$\begin{aligned} X(k) &= W_N^{-kN} \sum_{r=0}^{N-1} x(r) W_N^{kr} \\ &= \sum_{r=0}^{N-1} x(r) W_N^{-k(N-r)} \end{aligned} \quad (6.5)$$

Introduciamo ora, per comodità, la sequenza

$$y_k(n) = \sum_{r=0}^{N-1} x(r) W_N^{-k(n-r)} \quad (6.6)$$

Dalle (6.5) e (6.6) segue che

$$X(k) = y_k(n)|_{n=N}$$

La relazione (6.6) è chiaramente la convoluzione discreta della sequenza di durata finita $x(n)$, $0 \leq n \leq N-1$, con la sequenza W_N^{-kn} . Di conseguenza, $y_k(n)$ può essere vista come la risposta di un sistema con risposta all'impulso W_N^{-kn} a un ingresso $x(n)$. In particolare, $X(k)$ è il valore dell'uscita per $n = N$. Un sistema con risposta all'impulso W_N^{-kn} è rappresentato in fig. 6.1.

Poiché sia l'ingresso $x(n)$ che il coefficiente W_N^{-kn} sono complessi, il calcolo di ogni nuovo valore di $y_k(n)$ richiede quattro moltiplicazioni reali e quattro addizioni reali. Siccome poi occorre calcolare tutti i valori intermedi $y_k(1), y_k(2), \dots, y_k(N-1)$ per ottenere $y_k(N) = X(k)$, l'uso dello schema illustrato in fig. 6.1 richiede $4N$ moltiplicazioni reali e $4N$ addizioni reali per ricavare $X(k)$ per un particolare valore di k . Perciò questo

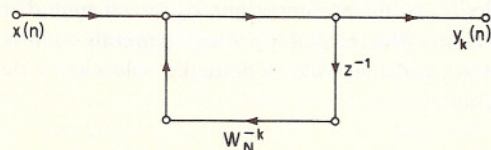


Fig. 6.1 Grafo di flusso di un sistema del primo ordine per il calcolo ricorsivo di $X(k)$.

schema è lievemente meno efficiente del metodo diretto. Notiamo però che il metodo di fig. 6.1 non richiede né il calcolo né la memorizzazione dei coefficienti W_N^{kn} , in quanto questi vengono calcolati attraverso il procedimento ricorsivo implicitamente contenuto nello schema di fig. 6.1.

È possibile mantenere questa semplificazione e ridurre contemporaneamente di un fattore 2 il numero delle moltiplicazioni. Per vederlo, notiamo che la funzione di trasferimento del sistema di fig. 6.1 è

$$H_k(z) = \frac{1}{1 - W_N^{-k} z^{-1}} \quad (6.7)$$

Moltiplicando il numeratore e il denominatore di $H_k(z)$ per il fattore $(1 - W_N^k z^{-1})$, si ottiene

$$\begin{aligned} H_k(z) &= \frac{1 - W_N^k z^{-1}}{(1 - W_N^{-k} z^{-1})(1 - W_N^k z^{-1})} \\ &= \frac{1 - W_N^k z^{-1}}{1 - 2 \cos((2\pi/N)k) z^{-1} + z^{-2}} \end{aligned} \quad (6.8)$$

La funzione di trasferimento (6.8) corrisponde al grafo di flusso di fig. 6.2.

Per realizzare i poli del sistema corrispondente alla (6.8) sono necessarie solo due moltiplicazioni, in quanto i coefficienti sono reali e il coefficiente (-1) non deve essere contato come moltiplicazione; il numero di addizioni è invece sempre quattro, come prima. Poiché basta portare il sistema in uno stato in cui sia possibile calcolare $y_k(N)$, la moltiplicazione complessa per $-W_N^k$, corrispondente allo zero, non deve essere eseguita ad ogni iterazione dell'equazione alle differenze, ma solo dopo l' N -ma. Quindi l'ammontare totale di operazioni è di $2N$ moltiplicazioni reali e $4N$ addizioni reali per i poli, più quattro moltiplicazioni reali e quattro addizioni reali per lo zero. Il peso complessivo dei calcoli è perciò di $2(N+2)$ moltiplicazioni reali e $4(N+1)$ addizioni reali, cioè circa la metà del numero di moltiplicazioni reali richieste dal metodo diretto. In questo schema, che è più efficiente del precedente, si conserva il vantaggio che gli unici coefficienti da calcolare e memorizzare sono $\cos((2\pi/N)k)$ e W_N^k , in quanto tutti i coefficienti W_N^{kn} sono ancora calcolati implicitamente nell'iterazione della formula ricorsiva rappresentata dalla fig. 6.2.

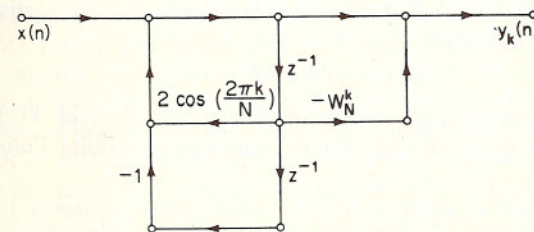


Fig. 6.2 Grafo di flusso di un sistema del secondo ordine per il calcolo ricorsivo di $X(k)$ (algoritmo di Goertzel).

Per illustrare un ulteriore vantaggio collegato all'uso di questo schema o rete, consideriamo il calcolo della trasformata z di $x(n)$ in punti coniugati sul circolo unitario, vale a dire il calcolo di $X(k)$ e di $X(N-k)$. È immediato verificare che la rete, analoga a quella di fig. 6.2, occorrente per calcolare $X(N-k)$ ha esattamente gli stessi poli di quella della fig. 6.2, mentre, per lo zero, il coefficiente è il complesso coniugato di quello di fig. 6.2 (v. probl. 6.1) Poiché l'operazione corrispondente allo zero viene eseguita solo all'ultima iterazione, le $2N$ moltiplicazioni e le $4N$ addizioni necessarie per i poli possono essere usate per ricavare due valori di DFT. Perciò, usando l'algoritmo di Goertzel, per il calcolo di tutti gli N punti della trasformata di Fourier discreta bastano circa N^2 moltiplicazioni e circa $2N^2$ addizioni. Tuttavia, il numero complessivo di operazioni è ancora proporzionale a N^2 , come per il calcolo diretto.

Nel metodo diretto o in quello di Goertzel non occorre ricavare tutti gli N diversi valori di $X(k)$. Anzi, possiamo in generale valutare $X(k)$ su M qualsiasi valori di k . In questo caso il peso totale dei calcoli è proporzionale a MN . Queste tecniche sono convenienti quando M è piccolo; sono però disponibili algoritmi più sofisticati per i quali il numero di operazioni è proporzionale a $N \log_2 N$ quando N è una potenza di 2. Quindi, quando M è minore di $\log_2 N$, il metodo di Goertzel o quello diretto possono davvero rappresentare la tecnica più efficiente, ma quando sono richiesti tutti gli N valori di $X(k)$, gli algoritmi che stiamo per prendere in esame sono circa $(N/\log_2 N)$ volte più efficienti del metodo diretto o di quello di Goertzel.

6.2 ALGORITMI DI FFT BASATI SULLA DECIMAZIONE NEL TEMPO

Per raggiungere il drastico aumento di efficienza cui abbiamo accennato, è necessario scomporre il calcolo della DFT in calcoli di DFT di dimensioni sempre più piccole. Nel fare questo sfruttiamo sia la simmetria che la periodicità dell'esponenziale complesso $W_N^{kn} = e^{-j(2\pi/N)kn}$. Gli algoritmi nei quali il procedimento di scomposizione si attua suddividendo la sequenza $x(n)$ in sottosequenze via via più piccole, si chiamano *algoritmi a decimazione nel tempo*. Il modo migliore di illustrare il principio della decimazione nel tempo è quello di considerare il caso particolare di N potenza intera di 2, cioè

$$N = 2^r$$

Poiché N è un intero pari, possiamo pensare di calcolare $X(k)$ dividendo $x(n)$ in due sequenze di $N/2$ punti³ ciascuna, costituite l'una dai punti

³ Nel discutere gli algoritmi di FFT, useremo indifferentemente i termini *campione* e *punto* col significato di singolo valore in una sequenza. Inoltre, indicheremo una sequenza di lunghezza N come una sequenza di N punti, e la DFT di una sequenza di lunghezza N come una DFT di (o su) N punti.

che hanno indice pari in $x(n)$, e l'altra da quelli con indice dispari. Se nell'espressione di $X(k)$

$$X(k) = \sum_{n=0}^{N-1} x(n)W_N^{nk}, \quad k = 0, 1, \dots, N-1 \quad (6.9)$$

scomponiamo $x(n)$ nei suoi punti con indice pari e con indice dispari, otteniamo

$$X(k) = \sum_{n \text{ pari}} x(n)W_N^{nk} + \sum_{n \text{ dispari}} x(n)W_N^{nk}$$

ovvero, con la sostituzione di variabili $n = 2r$ per n pari e $n = 2r + 1$ per n dispari,

$$\begin{aligned} X(k) &= \sum_{r=0}^{(N/2)-1} x(2r)W_N^{2rk} + \sum_{r=0}^{(N/2)-1} x(2r+1)W_N^{(2r+1)k} \\ &= \sum_{r=0}^{(N/2)-1} x(2r)(W_N^2)^{rk} + W_N^k \sum_{r=0}^{(N/2)-1} x(2r+1)(W_N^2)^{rk} \end{aligned} \quad (6.10)$$

Ma è $W_N^2 = W_{N/2}$ in quanto

$$W_N^2 = e^{-2j(2\pi/N)} = e^{-j2\pi/(N/2)} = W_{N/2}$$

Di conseguenza la (6.10) può essere riscritta come

$$\begin{aligned} X(k) &= \sum_{r=0}^{(N/2)-1} x(2r)W_{N/2}^{rk} + W_N^k \sum_{r=0}^{(N/2)-1} x(2r+1)W_{N/2}^{rk} \\ &= G(k) + W_N^k H(k) \end{aligned} \quad (6.11)$$

Si riconosce facilmente che ciascuna delle due sommatorie di questa espressione è una DFT di $N/2$ punti, essendo la prima sommatoria la DFT lunga $N/2$ dei punti con indice pari della sequenza originaria, ed essendo la seconda la DFT lunga $N/2$ dei punti con indice dispari della sequenza originaria. Anche se l'indice k può assumere N valori, $k = 0, 1, \dots, N-1$, occorre calcolare ogni somma solo per k tra 0 e $N/2 - 1$, in quanto sia $G(k)$ che $H(k)$ sono periodiche in k con periodo $N/2$. Dopo che sono state calcolate le due DFT corrispondenti alle due sommatorie della (6.11), esse devono essere combinate per ottenere la DFT su N punti, $X(k)$. La fig. 6.3 illustra il tipo di calcoli richiesto per ottenere $X(k)$ in base alla (6.11) per una sequenza di 8 punti, cioè per $N = 8$. In questa figura abbiamo usato le convenzioni relative ai grafi di flusso di segnale, introdotte nel cap. 4 per rappresentare le equazioni alle differenze [5, 7]. Così, i rami che entrano in un nodo si sommano per produrre la variabile del nodo. Quando manca l'indicazione del coefficiente di trasmissione di un ramo, il coefficiente è assunto unitario. Negli altri rami il coefficiente di trasmissione è una potenza intera di W_N . Quindi si vede dalla fig. 6.3 che vengono calcolate due DFT di 4 punti: $G(k)$ indica la DFT lunga quattro dei punti con indice pari e $H(k)$ indica la DFT lunga quattro dei punti con indice dispari. $X(0)$ si ottiene poi moltiplicando $H(0)$ per W_N^0 e sommando il ri-

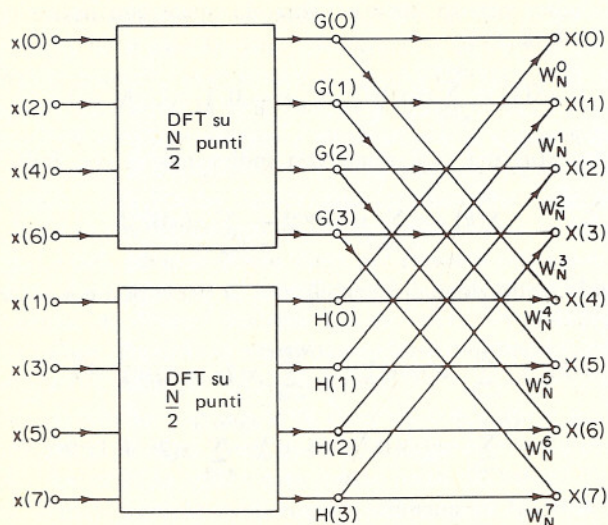


Fig. 6.3 Grafo di flusso per la scomposizione del calcolo di una DFT di N punti in due DFT di $N/2$ punti, con il metodo della decimazione nel tempo ($N=8$).

sultato a $G(0)$. $X(1)$ si ottiene moltiplicando $H(1)$ per W_N^1 e sommando il risultato a $G(1)$. Per $X(4)$ dovremmo moltiplicare $H(4)$ per W_N^4 e sommare il risultato a $G(4)$. Siccome però $G(k)$ e $H(k)$ sono entrambe periodiche in k con periodo 4, risulta $H(4) = H(0)$ e $G(4) = G(0)$. Perciò $X(4)$ si ottiene moltiplicando $H(0)$ per W_N^4 e sommando il risultato a $G(0)$.

Confrontiamo ora il numero di moltiplicazioni e addizioni richieste per il calcolo della DFT nel caso in cui si usi la struttura di calcolo indicata dalla (6.11) con il caso, già considerato prima, del calcolo diretto. Per quest'ultimo abbiamo visto che, se non si sfruttano le proprietà di simmetria, sono necessarie N^2 moltiplicazioni e addizioni complesse.⁴ A confronto, la (6.11) richiede il calcolo di due DFT su $N/2$ punti che a sua volta richiede $2(N/2)^2$ moltiplicazioni complesse e circa $2(N/2)^2$ addizioni complesse. Le due DFT su $N/2$ punti devono poi essere combinate come indicato nella (6.11), il che richiede altre N moltiplicazioni complesse, corrispondenti al prodotto della seconda sommatoria per W_N^k , ed altre N addizioni complesse, corrispondenti alla somma di quel prodotto con la prima sommatoria. Di conseguenza il calcolo della (6.11) per tutti i valori di k richiede $N + 2(N/2)^2$, cioè $N + N^2/2$, moltiplicazioni complesse e addizioni complesse. È facile verificare che, per $N > 2$, $N + N^2/2$ è minore di N^2 .

L'espressione (6.11) corrisponde a spezzare il calcolo originario su N punti in due calcoli su $N/2$ punti. Se $N/2$ è pari, come certamente avviene

⁴ Per semplicità, d'ora in poi assumeremo che N sia grande, in modo che $N - 1$ possa essere approssimato da N .

quando N è una potenza di 2, allora il calcolo di ciascuna DFT su $N/2$ punti nella (6.11) si può effettuare mediante il calcolo e la successiva combinazione di due DFT su $N/4$ punti. Pertanto $G(k)$ e $H(k)$ della (6.11) verrebbero calcolate come indicato qui appresso:

$$G(k) = \sum_{r=0}^{(N/2)-1} g(r) W_{N/2}^{rk} = \sum_{l=0}^{(N/4)-1} g(2l) W_{N/2}^{2lk} + \sum_{l=0}^{(N/4)-1} g(2l+1) W_{N/2}^{(2l+1)k}$$

oppure

$$G(k) = \sum_{l=0}^{(N/4)-1} g(2l) W_{N/4}^{lk} + W_{N/2}^k \sum_{l=0}^{(N/4)-1} g(2l+1) W_{N/4}^{lk} \quad (6.12)$$

Analogamente

$$H(k) = \sum_{l=0}^{(N/4)-1} h(2l) W_{N/4}^{lk} + W_{N/2}^k \sum_{l=0}^{(N/4)-1} h(2l+1) W_{N/4}^{lk} \quad (6.13)$$

Per il caso particolare della fig. 6.3, se ne deduce che, se le due DFT su 4 punti vengono calcolate seguendo le (6.12) e (6.13), i loro schemi di calcolo corrispondenti vengono ad essere come quelli indicati nella fig. 6.4. Introducendo tali schemi nel grafo di flusso della fig. 6.3, si ottiene il grafo di flusso completo della fig. 6.5. Notare che abbiamo fatto uso dell'identità $W_{N/2} = W_N^2$.

Per la DFT su 8 punti che abbiamo usato a mò di esempio, il calcolo si è ridotto a quello di due DFT su 2 punti. La DFT su 2 punti di, per esempio, $x(0)$ e $x(4)$, è schematizzata nella fig. 6.6. Inserendo tale schema nel grafo di flusso della fig. 6.5 si ottiene il grafo di flusso completo, mostrato nella fig. 6.7, per il calcolo della DFT su 8 punti.

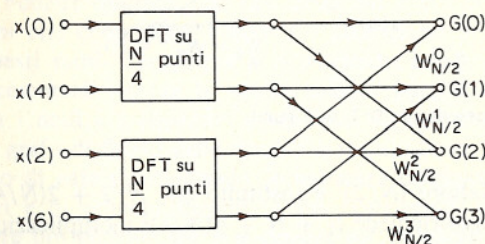


Fig. 6.4 Grafo di flusso per la scomposizione del calcolo di una DFT di $N/2$ punti in due DFT di $N/4$ punti, con il metodo della decimazione nel tempo ($N=8$).

Per il caso più generale in cui N è una potenza di 2 con esponente maggiore di 3, si procede scomponendo le trasformate su $N/4$ punti delle (6.12) e (6.13) in trasformate su $N/8$ punti, e così di seguito finché non si rimane con sole trasformate su 2 punti. Ciò richiede ν stadi di calcolo, dove $\nu = \log_2 N$. Prima si è visto che, con la originaria scomposizione di una trasformata su N punti in due trasformate su $N/2$ punti, il numero richiesto di moltiplicazioni e addizioni complesse era $N + 2(N/2)^2$. Quando le trasformate su $N/2$ punti vengono scomposte in trasformate su $N/4$

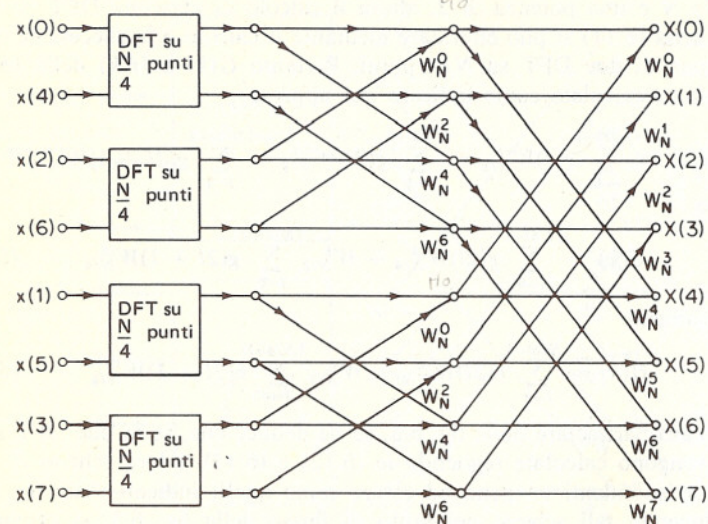


Fig. 6.5 Risultato che si ottiene introducendo nella fig. 6.3 gli schemi della fig. 6.4.

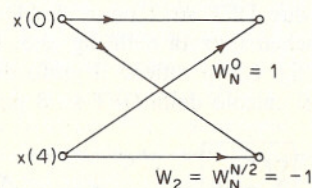


Fig. 6.6 Grafo di flusso di una DFT su 2 punti.

punti, allora il fattore $(N/2)^2$ è sostituito con $N/2 + 2(N/4)^2$, così che il calcolo complessivo richiede $N + N + 4(N/4)^2$ moltiplicazioni e addizioni complesse. Se $N = 2^v$, ciò può essere fatto al più $v = \log_2 N$ volte, e se ne conclude che, dopo aver iterato al massimo la scomposizione, il numero di moltiplicazioni e addizioni complesse diventa $N \log_2 N$.

Il grafo di flusso della fig. 6.7 mostra esplicitamente tutte le operazioni da fare. Se si contano i rami con coefficienti di trasmissione della forma W_N^k , si osserva che ogni stadio comporta N moltiplicazioni ed N addizioni complesse. Poiché gli stadi sono $\log_2 N$, se ne ricava, come prima, un totale di $N \log_2 N$ moltiplicazioni e addizioni complesse. È questo il sostanziale risparmio nei calcoli che prima avevamo indicato come possibile. Vedremo che si possono sfruttare la simmetria e la periodicità di W_N^k per ottenere ulteriori riduzioni nei calcoli.

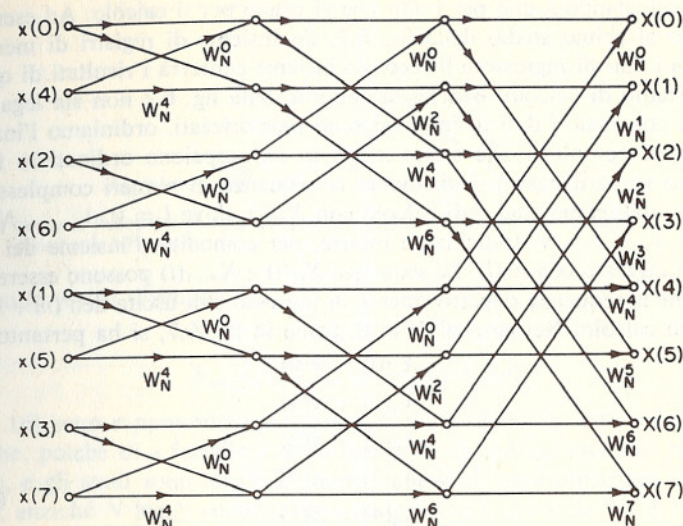


Fig. 6.7 Grafo di flusso per la scomposizione completa del calcolo di una DFT su 8 punti con il metodo della decimazione nel tempo.

6.2.1 Calcoli « sul posto »

Il grafo di flusso della fig. 6.7 descrive un algoritmo per il calcolo della trasformata di Fourier discreta. Ciò che conta nel grafo di flusso della fig. 6.7 sono i rami che connettono i nodi e i coefficienti di trasmissione di ciascuno di questi rami. Comunque si dispongano infatti i nodi, il grafo di flusso rappresenterà sempre lo stesso calcolo purché si conservino le connessioni tra i nodi e i rispettivi coefficienti di trasmissione. La forma particolare del grafo di flusso della fig. 6.7 è venuta fuori perché si è derivato l'algoritmo di calcolo separando la sequenza originaria nei campioni di numero pari e di numero dispari, e poi costruendo allo stesso modo sequenze sempre più piccole. Un interessante sottoprodotto di questa derivazione è che il grafo di flusso della fig. 6.7, oltre a descrivere una procedura efficiente per il calcolo della trasformata di Fourier discreta, suggerisce anche un modo utile di memorizzare i dati originari e i risultati dei calcoli negli stadi intermedi.

Per rendercene conto cominciamo con l'osservare che, stando alla fig. 6.7, ogni stadio di calcolo prende, per così dire, N numeri complessi e li trasforma in un altro insieme di N numeri complessi. Questo processo si ripete un numero di volte pari a $v = \log_2 N$, dando luogo infine alla trasformata di Fourier discreta desiderata. Per realizzare effettivamente i calcoli schematizzati nella fig. 6.7 possiamo immaginare di usare due insiemi di registri di memoria (complessi), uno per la sequenza di valori che si

stanno calcolando e uno per i dati che si usano per il calcolo. Ad esempio, riguardo al primo stadio della fig. 6.7, un insieme di registri di memoria conterrà i dati di ingresso e il secondo insieme conterrà i risultati di questo primo stadio di calcolo. Benché la validità della fig. 6.7 non sia legata all'ordine col quale i dati di ingresso sono memorizzati, ordiniamo l'insieme dei numeri complessi allo stesso modo in cui appaiono ordinati in figura (dall'alto verso il basso). Indichiamo la sequenza di numeri complessi che risultano dall' m .mo stadio di calcolo con $X_m(l)$, dove $l = 0, 1, \dots, N-1$, ed $m = 1, 2, \dots, v$. Indichiamo inoltre, per comodità, l'insieme dei campioni di ingresso con $X_0(l)$. Le sequenze $X_m(l)$ e $X_{m+1}(l)$ possono essere pensate come le sequenze rispettivamente di ingresso e di uscita dell' $(m+1)$.mo stadio di calcolo. Nel caso di $N = 8$, come in fig. 6.7, si ha pertanto

$$\begin{aligned} X_0(0) &= x(0) \\ X_0(1) &= x(4) \\ X_0(2) &= x(2) \\ X_0(3) &= x(6) \\ X_0(4) &= x(1) \\ X_0(5) &= x(5) \\ X_0(6) &= x(3) \\ X_0(7) &= x(7) \end{aligned} \quad (6.14)$$

Usando queste notazioni si verifica facilmente che il calcolo base del grafo di flusso della fig. 6.7 è quello mostrato nella fig. 6.8.

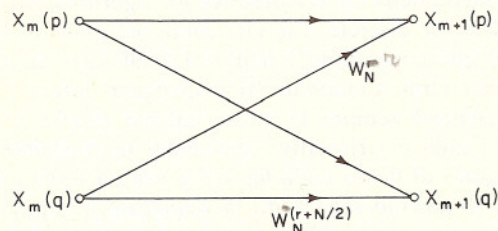


Fig. 6.8 Grafo di flusso del calcolo base (« farfalla ») della fig. 6.7.

Le equazioni rappresentate da questo grafo di flusso sono della forma

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + W_N^r X_m(q) \\ X_{m+1}(q) &= X_m(p) - W_N^{r+N/2} X_m(q) \end{aligned} \quad (6.15)$$

A causa della sagoma del grafo di flusso della fig. 6.8 questo calcolo viene comunemente chiamato calcolo a (o della) *farfalla*.

Le espressioni (6.15) suggeriscono un mezzo per ridurre di un fattore due il numero delle moltiplicazioni complesse. Per rendercene conto cominciamo con l'osservare che

$$W_N^{N/2} = e^{-j(2\pi/N) \cdot N/2} = e^{-j\pi} = -1$$

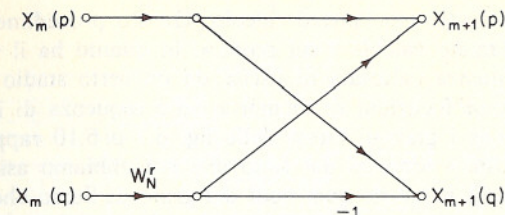


Fig. 6.9 Grafo di flusso del calcolo semplificato della farfalla richiedente soltanto una moltiplicazione complessa.

per cui le (6.15) diventano

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + W_N^r X_m(q) \\ X_{m+1}(q) &= X_m(p) - W_N^r X_m(q) \end{aligned} \quad (6.16)$$

Le (6.16) sono rappresentate nel grafo di flusso della fig. 6.9. Osserviamo ora che, poiché di « farfalle » della forma di fig. 6.9 ne esistono $N/2$ per stadio, e gli stadi sono $\log_2 N$, il numero totale di moltiplicazioni è $(N/2) \log_2 N$, anziché $N \log_2 N$ come sembra dalla fig. 6.7. Se nella fig. 6.7 le farfalle della forma di fig. 6.8 vengono sostituite dal grafo di flusso base della fig. 6.9, otteniamo il grafo di flusso della fig. 6.10. Il numero totale di moltiplicazioni complesse necessarie per il calcolo risulta anche evidente da tale figura in base a una semplice ispezione.

Nella (6.16) i parametri p, q ed r variano da stadio a stadio in modo rapidamente deducibile dalla fig. 6.10 e dalle (6.10), (6.12), (6.13), ecc. È inoltre chiaro dalle fig. 6.9 e 6.10 che per il calcolo degli elementi in posizione p e q della $(m+1)$.ma sequenza di numeri complessi sono necessari gli elementi p e q della sequenza m .ma. Pertanto, se $X_{m+1}(p)$ e $X_{m+1}(q)$ vengono memorizzati negli stessi registri di memoria usati per $X_m(p)$ e $X_m(q)$, sarà necessario fisicamente un solo insieme di N registri di memoria

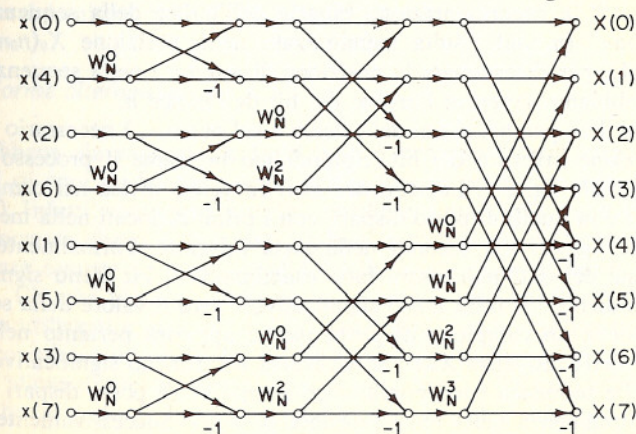


Fig. 6.10 Grafo di flusso di una DFT su 8 punti facente uso del calcolo della farfalla di fig. 6.9.

(complessi) per realizzare l'intero calcolo. Questo procedimento viene comunemente chiamato calcolo « sul posto », in quanto ha il vantaggio che ogni nuova sequenza calcolata in uscita da un certo stadio viene memorizzata nelle stesse locazioni di memoria della sequenza di ingresso originaria. Il fatto che i grafi di flusso delle fig. 6.7 o 6.10 rappresentino dei calcoli « sul posto » dipende dal fatto che noi abbiamo associato con la stessa locazione di memoria quei nodi del grafo di flusso che stanno sulla stessa linea orizzontale, e che il calcolo fra due sequenze (di ingresso e uscita a un certo stadio) consiste di un calcolo a farfalla nel quale i nodi di ingresso e i nodi di uscita sono orizzontalmente adiacenti.

Affinché il calcolo possa essere fatto « sul posto » osserviamo che con i grafi di flusso di fig. 6.7 (o 6.10) i dati di ingresso devono essere memorizzati in ordine non sequenziale. In effetti l'ordine col quale i dati di ingresso sono memorizzati è *a disposizione invertita dei bit*. Per vedere che cosa si intende con questa terminologia, cominciamo con l'osservare che, per il grafo di flusso su 8 punti fin qui discusso, sono necessarie tre cifre binarie per dare un indice a ciascun dato. Se ora scriviamo gli indici della (6.14) in forma binaria otteniamo l'insieme di relazioni

$$\begin{aligned} X_0(000) &= x(000) \\ X_0(001) &= x(100) \\ X_0(010) &= x(010) \\ X_0(011) &= x(110) \\ X_0(100) &= x(001) \\ X_0(101) &= x(101) \\ X_0(110) &= x(011) \\ X_0(111) &= x(111) \end{aligned} \quad (6.17)$$

Se $(n_2 n_1 n_0)$ è la rappresentazione binaria dell'indice della sequenza $x(n)$, il campione $x(n_2 n_1 n_0)$ risulta memorizzato nella posizione $X_0(n_0 n_1 n_2)$. In altre parole, per determinare la posizione di $x(n_2 n_1 n_0)$ nella sequenza di ingresso, dobbiamo invertire l'ordine dei bit dell'indice n .

Per vedere perché, ai fini del calcolo « sul posto », è necessario l'ordine a disposizione invertita dei bit, riprendiamo in esame il processo che ha dato luogo alla fig. 6.7. La sequenza $x(n)$ fu prima divisa nei campioni di posto pari e in quelli di posto dispari, con i primi collocati nella metà superiore della fig. 6.3 e i secondi nella metà inferiore. Formalmente questa separazione dei dati può essere fatta esaminando il bit meno significativo (n_0) nell'indice n . Se il bit meno significativo è zero il valore della sequenza corrisponde a un campione di posto pari e apparirà pertanto nella metà superiore della sequenza $X_0(l)$; se viceversa il bit meno significativo è 1, il valore della sequenza corrisponde a un campione di posto dispari e apparirà di conseguenza nella metà inferiore di $X_0(l)$. Successivamente le due sottosequenze pari e dispari vengono separate nelle loro parti pari e dispari, e ciò può essere fatto esaminando il secondo bit meno significativo nell'in-

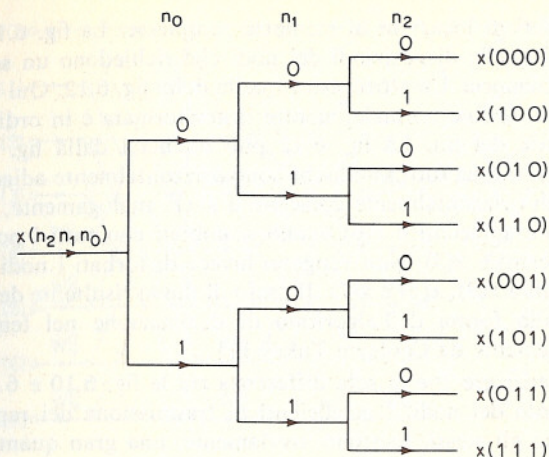


Fig. 6.11 Diagramma ad albero raffigurante l'ordinamento dei campioni con la disposizione invertita dei bit.

dice dei dati. Considerando per prima la sottosequenza costituita dai campioni di posto pari, se il secondo bit meno significativo è zero, il valore della sequenza è un campione di posto pari all'interno di questa sottosequenza; se viceversa il secondo bit meno significativo è 1, allora il valore della sequenza ha posto dispari all'interno della stessa. Lo stesso procedimento si applica alla sottosequenza costituita dai campioni di posto dispari. Tale procedimento viene ripetuto finché si ottengono N sottosequenze di lunghezza 1. Questa separazione dei campioni in sottosequenze con indici pari e dispari è raffigurata nel diagramma ad albero della fig. 6.11.

Pertanto la necessità dell'ordinamento con la disposizione invertita dei bit è una conseguenza del modo in cui il calcolo della DFT è scomposto in successivi calcoli di DFT via via più piccole.

6.2.2 Forme alternative

Sebbene sia ragionevole farlo, non è però necessario memorizzare i risultati di ogni stadio del calcolo nell'ordine in cui i nodi appaiono nella fig. 6.10. Infatti, comunque vengano riordinati i nodi della fig. 6.10, il risultato, ammesso che non vengano cambiati i coefficienti di trasmissione dei rami, sarà sempre un calcolo valido della trasformata di Fourier discreta di $x(n)$. Ciò che viene a cambiare in tal caso è solo l'ordine in cui i dati vengono utilizzati e memorizzati. Se associamo i nodi con locazioni di memoria complesse in modo che il loro ordine sia associato con gli indici di queste, allora dovrebbe essere chiaro, sulla base delle osservazioni precedenti, che un grafo di flusso corrisponde a un calcolo « sul posto » solo se il riordinamento dei nodi è tale che i nodi di ingresso e di uscita per ogni « farfalla » sono orizzontalmente adiacenti. Se non è così, occorrono due

insiemi completi di locazioni di memoria complesse. La fig. 6.10 è, ovviamente, una di quelle disposizioni dei nodi che richiedono un solo insieme di registri di memoria. Un altro caso è quello della fig. 6.12. Qui la sequenza di ingresso è in ordine normale, mentre la trasformata è in ordine a disposizione invertita dei bit. La fig. 6.12 può ottenersi dalla fig. 6.10 scambiando in quest'ultima tutti i nodi che sono orizzontalmente adiacenti a $x(4)$ con tutti i nodi orizzontalmente adiacenti a $x(1)$; analogamente, tutti i nodi orizzontalmente adiacenti a $x(6)$ vanno scambiati con tutti i nodi orizzontalmente adiacenti a $x(3)$. Non vengono invece disturbati i nodi orizzontalmente adiacenti a $x(2)$, $x(5)$ e $x(7)$. Il grafo di flusso risultante della fig. 6.12 corrisponde alla forma dell'algoritmo di decimazione nel tempo sviluppato originariamente da Cooley e Tukey [1].

Si può verificare che la sola differenza tra le fig. 6.10 e 6.12 consiste nell'ordinamento dei nodi. I coefficienti di trasmissione dei rami (potenze di W_N) restano gli stessi. Esistono, ovviamente, una gran quantità di ordinamenti possibili; la maggior parte, tuttavia, non è gran che significativa dal punto di vista dei calcoli. Un'altra possibilità utile è quella in cui i nodi sono ordinati in modo che tanto la sequenza di ingresso che quella di uscita si presentano nell'ordine normale. Un grafo di flusso di questo tipo è mostrato nella fig. 6.13. In questo caso, tuttavia, il calcolo non può essere fatto « sul posto ». Pertanto per eseguire i calcoli indicati nella fig. 6.13 sarebbero necessari due insiemi di registri complessi di lunghezza N .

Per realizzare i calcoli rappresentati nelle fig. 6.10, 6.12 e 6.13, è chiaramente necessario poter accedere agli elementi delle sequenze intermedie, $X_m(l)$, in ordine non sequenziale. Pertanto, per ottenere maggiori velocità di calcolo, i numeri complessi devono essere memorizzati in memorie ad accesso casuale. Per esempio, riferendoci alla fig. 6.10, osserviamo che nel calcolo delle uscite del primo stadio gli ingressi a ciascuna « far-

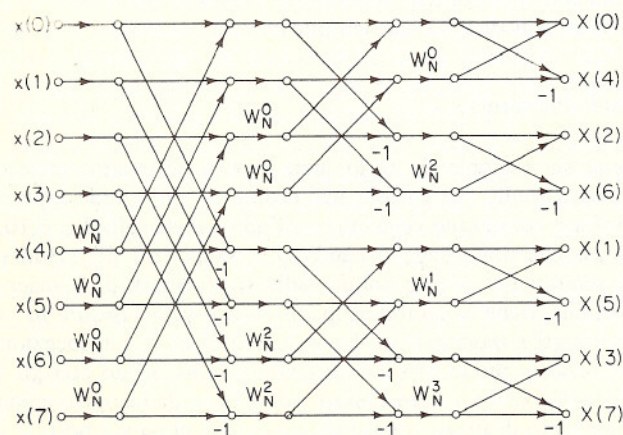


Fig. 6.12 Riordinamento dei nodi della fig. 6.10 con l'ingresso in ordine normale e l'uscita in ordine a disposizione invertita dei bit.

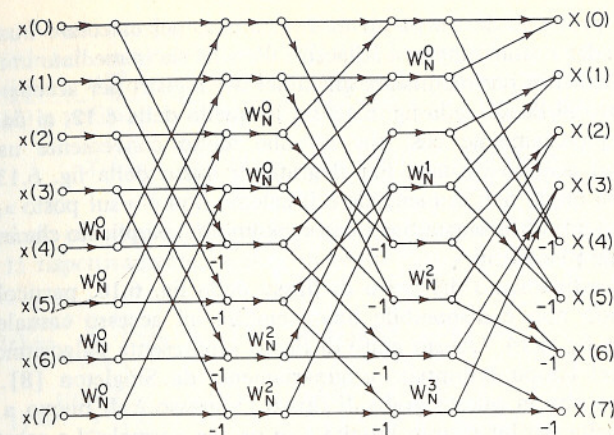


Fig. 6.13 Riordinamento dei nodi della fig. 6.10 con tanto l'ingresso che l'uscita in ordine normale.

falla » sono variabili di nodi adiacenti che si possono pensare immagazzinate in locazioni di memoria adiacenti. Nel calcolo delle uscite del secondo stadio, invece, gli ingressi a ciascuna « farfalla » sono separati da due posizioni di memoria, mentre per il calcolo delle uscite del terzo stadio gli ingressi a ciascuna « farfalla » sono separati da quattro posizioni di memoria. Se N fosse maggiore di 8, la separazione tra gli ingressi delle farfalle sarebbe 8 per il quarto stadio, 16 per il quinto, ecc. La separazione per l'ultimo stadio (v.mo) sarebbe $N/2$.

Nella fig. 6.12 la situazione è simile, in quanto nel calcolare l'uscita dal primo stadio si devono usare ingressi (in questo caso dati) separati di 4, nel calcolare l'uscita del secondo stadio si devono usare ingressi (le

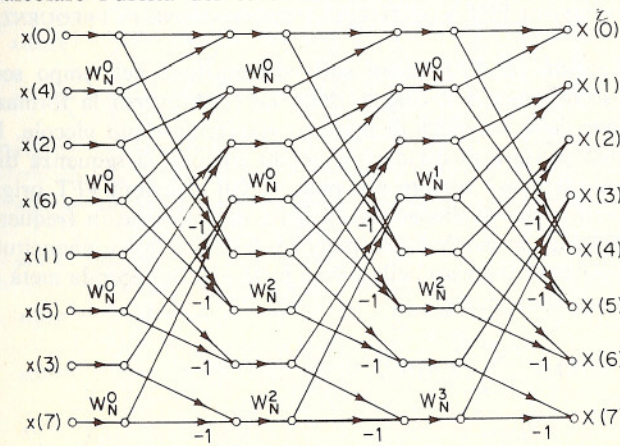


Fig. 6.14 Riordinamento della fig. 6.10 caratterizzato dalla stessa geometria per ogni stadio, per cui è possibile accedere ai dati e memorizzarli sequenzialmente.

uscite del primo stadio) separati di 2 e, infine, nel calcolare l'uscita dell'ultimo stadio si usano ingressi adiacenti. Benché sia immediato immaginare semplici algoritmi per modificare gli indici dei registri per accedere ai dati sia nel grafo di flusso della fig. 6.10 sia in quello della 6.12, ai dati non si accede sequenzialmente così che sarebbe molto conveniente usare una memoria ad accesso casuale. Per il grafo di flusso della fig. 6.13 ai dati si accede in modo non sequenziale e il calcolo non è « sul posto », per cui uno schema di indirizzamento dei dati è molto più complicato che in ognuno dei due casi precedenti.

Un riordinamento del grafo di flusso della fig. 6.10, particolarmente utile quando non è disponibile una memoria ad accesso casuale, è mostrato nella fig. 6.14. Questo grafo di flusso rappresenta l'algoritmo a decimazione nel tempo sviluppato originariamente da Singleton [8]. Si noti innanzitutto che in questo grafo di flusso l'ingresso è di nuovo a disposizione invertita dei bit mentre l'uscita è in ordine normale. La caratteristica importante di questo grafo di flusso è che la geometria è la stessa per ogni stadio; da stadio a stadio cambiano solo i coefficienti di trasmissione dei rami. Ciò rende possibile accedere ai dati in modo sequenziale. Supponiamo infatti di disporre di quattro unità nastro magnetico (o di quattro aree sequenziali di memoria a disco) e supponiamo che la prima metà dei dati di ingresso (in ordine a disposizione invertita dei bit) sia immagazzinata su un nastro e l'altra metà su un altro nastro. Allora si può accedere sequenzialmente ai dati sui nastri 1 e 2 e scrivere sequenzialmente i risultati sui nastri 3 e 4, con la prima metà della nuova sequenza sul nastro 3 e la seconda metà sul nastro 4. Quindi, al nuovo stadio di calcolo, i nastri 3 e 4 diventano gli ingressi e le uscite vengono scritte sui nastri 1 e 2. Ciò si ripete per ognuno dei ν stadi.

6.3 ALGORITMI DI FFT BASATI SULLA DECIMAZIONE IN FREQUENZA

Gli algoritmi di FFT basati sulla decimazione nel tempo sono stati sviluppati scomponendo il calcolo della DFT attraverso la formazione di sottosequenze della sequenza di ingresso $x(n)$ sempre più piccole. In alternativa si può pensare di dividere in modo analogo la sequenza di uscita, $X(k)$, in sottosequenze sempre più piccole. Gli algoritmi FFT originati da questo procedimento si dicono basati sulla decimazione in frequenza. Per derivarli, nel caso in cui N è una potenza di 2, possiamo innanzitutto dividere la sequenza di ingresso nella prima metà e nella seconda metà dei suoi punti, in modo da scrivere

$$X(k) = \sum_{n=0}^{(N/2)-1} x(n) W_N^{nk} + \sum_{n=N/2}^{N-1} x(n) W_N^{nk}$$

oppure

$$X(k) = \sum_{n=0}^{(N/2)-1} x(n) W_N^{nk} + W_N^{(N/2)k} \sum_{n=0}^{(N/2)-1} x\left(n + \frac{N}{2}\right) W_N^{nk} \quad (6.18)$$

È importante osservare che, pur contenendo la (6.18) due sommatorie su $N/2$, ciascuna di esse non è una DFT su $N/2$ punti, in quanto nelle sommatorie appare W_N^{nk} e non $W_{(N/2)}^{nk}$. Mettendo insieme le due sommatorie nella (6.18) e utilizzando il fatto che $W_N^{(N/2)k} = (-1)^k$, si ottiene

$$X(k) = \sum_{n=0}^{(N/2)-1} \left[x(n) + (-1)^k x\left(n + \frac{N}{2}\right) \right] W_N^{nk} \quad (6.19)$$

Considerando ora separatamente k pari e k dispari, indicando con $X(2r)$ e $X(2r+1)$ rispettivamente i valori di posto pari a quelli di posto dispari, avremo:

$$X(2r) = \sum_{n=0}^{(N/2)-1} \left[x(n) + x\left(n + \frac{N}{2}\right) \right] W_N^{2rn} = P_1(k) \quad (6.20)$$

$$X(2r+1) = \sum_{n=0}^{(N/2)-1} \left[x(n) - x\left(n + \frac{N}{2}\right) \right] W_N^{2rn} = D_1(k) \quad (6.21)$$

$r = 0, 1, \dots, (N/2 - 1)$

In queste due espressioni possiamo ora riconoscere due DFT su $N/2$ punti; nel caso della (6.20) si tratta della DFT della somma della prima metà e della seconda metà della sequenza di ingresso, e nel caso della (6.21) si tratta della DFT del prodotto di W_N^n con la differenza fra la prima metà e la seconda metà della sequenza di ingresso. Differentemente dal caso della (6.19), le sommatorie nelle (6.20) e (6.21) corrispondono a delle DFT su $N/2$ punti in quanto

$$W_N^{2rn} = W_{N/2}^{rn} \quad P_1(n)$$

Pertanto, sulla base delle (6.20) e (6.21), ponendo $g(n) = x(n) + x(n + N/2)$ e $h(n) = x(n) - x(n + N/2)$, la DFT può essere calcolata formando innanzitutto le sequenze $g(n)$ ed $h(n)$, poi calcolando $h(n) W_N^n$ e infine

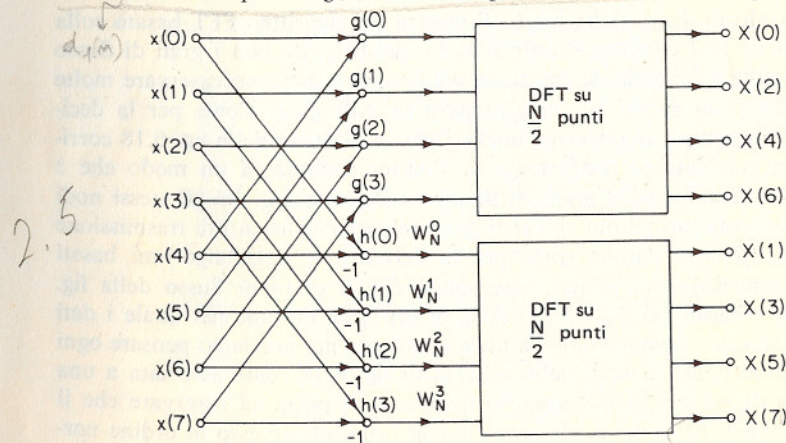


Fig. 6.15 Grafo di flusso per la scomposizione del calcolo di una DFT di N punti in due DFT di $N/2$ punti, con il metodo della decimazione in frequenza ($N=8$).

calcolando le DFT su $N/2$ punti di queste due sequenze, ricavandone rispettivamente i valori di uscita di posto pari e quelli di posto dispari. Il procedimento suggerito dalle (6.20) e (6.21) è illustrato nella fig. 6.15 nel caso di una DFT su 8 punti.

Procedendo in modo simile a quello usato per derivare gli algoritmi basati sulla decimazione nel tempo, notiamo che, essendo N una potenza di 2, $N/2$ è pari e di conseguenza le DFT su $N/2$ punti possono essere effettuate calcolando separatamente per ciascuna di esse i valori di uscita di posto pari e quelli di posto dispari. Come nel caso della scomposizione precedente che ha portato alle (6.20) e (6.21), ciò si realizza combinando la prima metà e la seconda metà dei punti di ingresso per ognuna delle DFT su $N/2$ punti e calcolando quindi delle DFT su $N/4$ punti. Lo schema di flusso risultante nel caso dell'esempio su 8 punti è mostrato nella fig. 6.16. In questo particolare caso il calcolo si è così ridotto a quello di alcune DFT su 2 punti, le quali, come abbiamo visto in precedenza, si calcolano sommando e sottraendo i punti di ingresso. Pertanto le DFT su 2 punti della fig. 6.16 possono essere sostituite con lo schema di calcolo mostrato nella fig. 6.17, così che il calcolo completo della DFT su 8 punti diventa quello mostrato nella fig. 6.18.

Se si contano le operazioni aritmetiche nella fig. 6.18, e si generalizza al caso $N = 2^n$, si vede che il calcolo della DFT richiede $N/2 \log_2 N$ moltiplicazioni complesse e $N \log_2 N$ addizioni complesse. Pertanto il numero complessivo di operazioni è lo stesso per gli algoritmi basati sulla decimazione in frequenza e sulla decimazione nel tempo.

6.3.1 Calcoli « sul posto »

Il grafo di flusso della fig. 6.18 mostra un algoritmo FFT basato sulla decimazione in frequenza. Confrontando questo grafo con i grafi di flusso derivati sulla base della decimazione nel tempo, si possono osservare molte somiglianze, ma anche un certo numero di differenze. Come per la decimazione nel tempo, ovviamente anche il grafo di flusso della fig. 6.18 corrisponde a calcolare la trasformata di Fourier discreta in un modo che è indipendente da come il grafo di flusso è disegnato, purché gli stessi nodi rimangano collegati gli uni agli altri con i relativi coefficienti di trasmissione sui rami. In altre parole, come per la derivazione degli algoritmi basati sulla decimazione nel tempo, osserviamo che il grafo di flusso della fig. 6.18 non è basato su alcuna ipotesi a priori circa l'ordine nel quale i dati di ingresso sono memorizzati. Tuttavia anche ora noi possiamo pensare ogni sequenza verticale di nodi nello schema di fig. 6.18 come associata a una sequenza di registri di memoria, la qual cosa ci porta ad osservare che il grafo di flusso della fig. 6.18 comincia con i dati di ingresso in ordine normale e fornisce i valori di uscita in ordine a disposizione invertita dei bit. Notiamo ancora che il calcolo base è anche qui del tipo « a farfalla »,

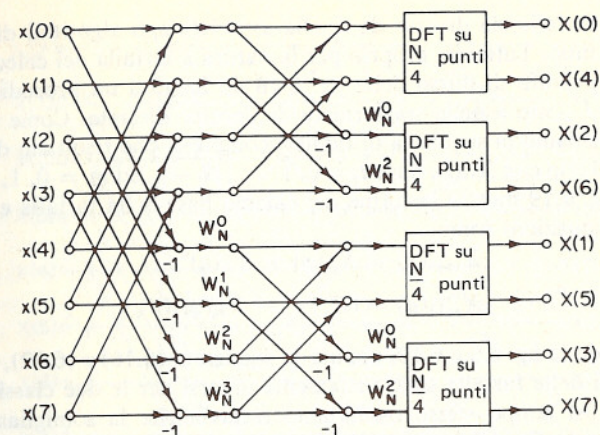


Fig. 6.16 Grafo di flusso per la scomposizione del calcolo di una DFT di 8 punti nel calcolo di 4 DFT di 2 punti, con il metodo della decimazione in frequenza.

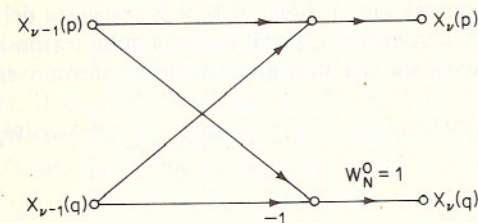


Fig. 6.17 Grafo di flusso di una DFT su 2 punti del tipo richiesto all'ultimo stadio di calcolo di una scomposizione basata sulla decimazione in frequenza.

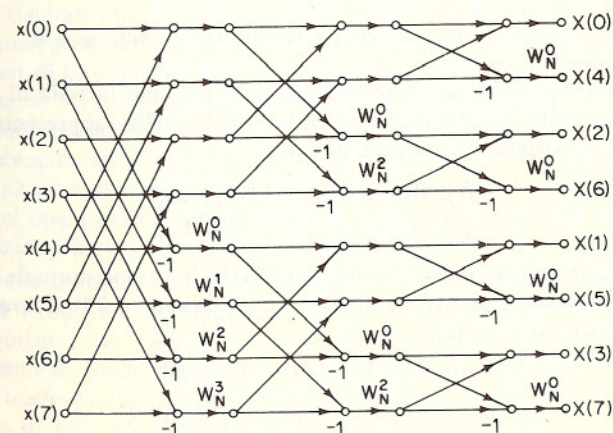


Fig. 6.18 Grafo di flusso per la scomposizione completa di una DFT di 8 punti con il metodo della decimazione in frequenza.

sebbene la farfalla sia diversa da quella associata agli algoritmi di decimazione nel tempo. Tuttavia, proprio per la natura a farfalla dei calcoli, osserviamo che il grafo di flusso della fig. 6.18 dà luogo a un procedimento di calcolo « sul posto » della trasformata di Fourier discreta. Come in precedenza, indichiamo la sequenza di numeri complessi che risultano dall' m -esimo stadio di calcolo con $X_m(l)$, dove $l = 0, 1, \dots, N-1$, ed $m = 0, 1, 2, \dots, v$. Allora la fig. 6.19 mostra la forma del calcolo base della farfalla e le corrispondenti equazioni sono

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + X_m(q) \\ X_{m+1}(q) &= (X_m(p) - X_m(q))W_N^r \end{aligned} \quad (6.22)$$

Confrontando le fig. 6.9 e 6.19, oppure le equazioni (6.16) e (6.22), vediamo che i calcoli delle farfalle sono nettamente diversi per le due classi di algoritmi FFT. Al tempo stesso osserviamo tuttavia che la somiglianza tra i calcoli o tra le fig. 6.10 e 6.18 è notevolissima. Più precisamente, osserviamo che la fig. 6.18 può essere ottenuta dalla fig. 6.10 invertendo la direzione del flusso dei segnali e scambiando ingressi e uscite. In altre parole, con la terminologia del cap. 4, la fig. 6.18 è la trasposta del grafo di flusso della fig. 6.10 e di conseguenza, per il teorema della trasposizione, le caratteristiche ingresso-uscita dei due grafi di flusso devono essere le stesse.

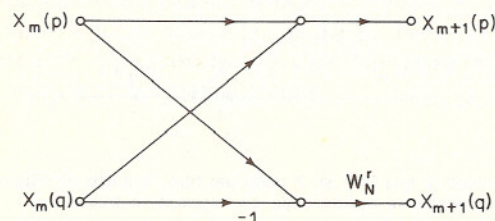


Fig. 6.19 Grafo di flusso di un tipico calcolo a farfalla richiesto nella fig. 6.18.

Per vedere tutto ciò in un altro modo si consideri una farfalla di un algoritmo basato sulla decimazione nel tempo come quella rappresentata nella fig. 6.9. Le corrispondenti equazioni sono

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + W_N^r X_m(q) \\ X_{m+1}(q) &= X_m(p) - W_N^r X_m(q) \end{aligned} \quad (6.23)$$

Se supponiamo di cominciare con le uscite $X(k)$ in ordine normale come in fig. 6.10, possiamo invertire i calcoli delle equazioni (6.23) ricavando X_m in funzione di X_{m+1} , cioè:

$$\begin{aligned} X_m(p) &= \frac{1}{2}(X_{m+1}(p) + X_{m+1}(q)) \\ X_m(q) &= \frac{1}{2}(X_{m+1}(p) - X_{m+1}(q))W_N^{-r} \end{aligned} \quad (6.24)$$

Pertanto, poiché

$$X_v(k) = X(k)$$

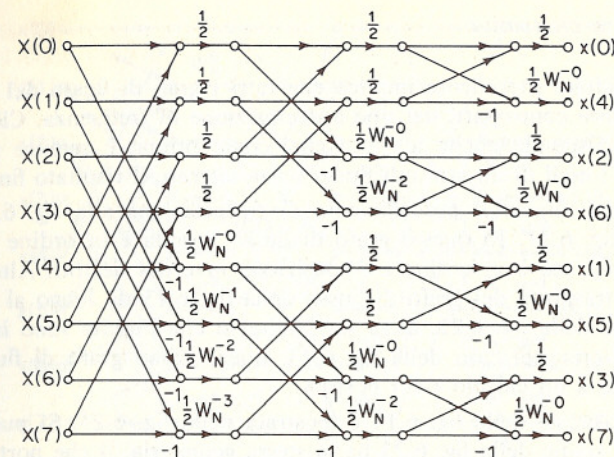


Fig. 6.20 Grafo di flusso di un calcolo di DFT inversa ottenuto invertendo i calcoli a farfalla di fig. 6.10.

e $X_0(k)$ coincide $x(n)$ a parte l'ordine, che è a disposizione invertita dei bit, possiamo calcolare $x(n)$ nell'ordine suddetto applicando ripetutamente le equazioni (6.24). Il grafo di flusso per $N = 8$ è mostrato nella fig. 6.20.

La fig. 6.20 descrive un algoritmo di trasformata di Fourier veloce inversa (IFFT). Osserviamo ora che la trasformata di Fourier discreta inversa (IDFT) è

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}$$

per cui per calcolare la IDFT si può usare un algoritmo di FFT se si divide il risultato per N e se si usano potenze di W_N^{-1} invece di potenze di W_N . Analogamente, un algoritmo di IFFT può essere usato per calcolare la DFT pur di moltiplicare l'uscita per N e di usare potenze di W_N al posto di potenze di W_N^{-1} . Ne segue che il grafo di flusso della fig. 6.20, rappresentante un algoritmo di IFFT, può essere fatto diventare un algoritmo di FFT cambiando semplicemente $(1/2)W_N^{-r}$ con W_N^r , in quanto eliminare il fattore $1/2$ ad ogni stadio equivale a moltiplicare l'uscita per N . Se nella fig. 6.20 si opera questo cambiamento e si pone all'ingresso $x(n)$ in ordine a disposizione invertita dei bit, si ottiene la fig. 6.18.

Si vede pertanto che ad ogni algoritmo di FFT basato sulla decimazione nel tempo corrisponde un algoritmo di trasformazione inversa che è un algoritmo basato sulla decimazione in frequenza. Oppure, essendo gli algoritmi di trasformazione inversa in semplice relazione con gli algoritmi di trasformazione diretta, si può affermare in generale che per ogni algoritmo di FFT a decimazione nel tempo esiste un algoritmo di FFT a decimazione in frequenza il quale corrisponde a scambiare ingressi e uscite e ad invertire la direzione di tutte le frecce nel grafo di flusso.

6.3.2 Forme alternative

Il risultato precedente implica che tutti i grafi di flusso del par. 6.2 hanno le loro controparti del tipo a decimazione in frequenza. Ciò, ovviamente, corrisponde anche al fatto che, come prima, è sempre possibile riordinare i nodi di un grafo di flusso senza alterare il risultato finale.

Se applichiamo il procedimento di trasposizione alla fig. 6.12 otteniamo la fig. 6.21. In questo grafo di flusso l'uscita è in ordine normale mentre l'ingresso è in ordine a disposizione invertita dei bit. Alternativamente, la trasposta del grafo di flusso della fig. 6.13 dà luogo al grafo di flusso della fig. 6.22, dove sia l'ingresso che l'uscita sono in ordine normale. Come nel caso della fig. 6.13, quest'ultimo grafo di flusso non corrisponde a un calcolo « sul posto ».

La trasposta della fig. 6.14 è mostrata nella fig. 6.23. Si può notare come ogni stadio della fig. 6.23 ha la stessa geometria, il che porta a quei vantaggi già discussi in precedenza per lo svolgimento dei calcoli mediante memorizzazione sequenziale dei dati.

6.4 ALGORITMI DI FFT PER N NUMERO COMPOSTO

La discussione precedente ha illustrato i principi fondamentali della decimazione nel tempo e della decimazione in frequenza per l'importante caso particolare in cui N è una potenza di 2, cioè $N = 2^v$. Più in generale, il calcolo efficiente della trasformata di Fourier discreta è legato alla rappresentazione di N come prodotto di fattori [1, 7, 9, 10]. Si supponga pertanto

$$N = p_1 p_2 \dots p_r \quad (6.25)$$

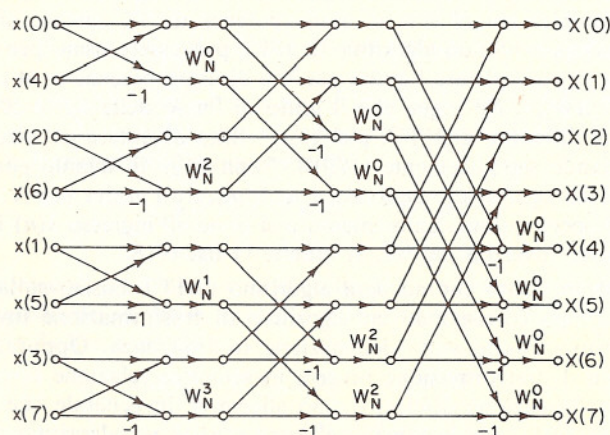


Fig. 6.21 Grafo di flusso di un algoritmo di DFT a decimazione in frequenza ottenuto dalla fig. 6.18. L'ingresso è in ordine a disposizione invertita dei bit e l'uscita in ordine normale. (Trasposta della fig. 6.12).

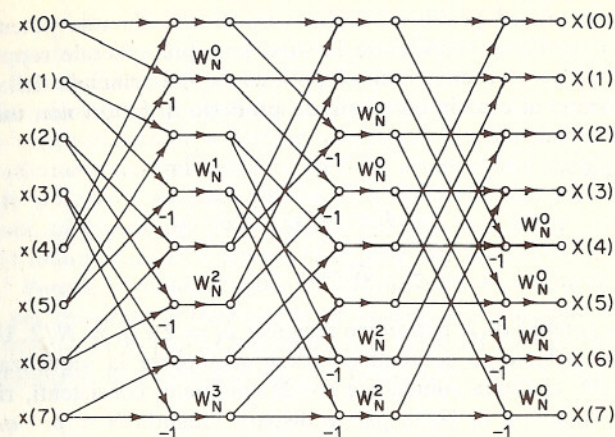


Fig. 6.22 Riordinamento della fig. 6.18 con tanto l'ingresso che l'uscita in ordine normale. (Trasposta della fig. 6.13).

Come abbiamo visto per N potenza di 2 (nel qual caso tutti i fattori possono essere presi uguali a 2), una tale scomposizione conduce ad un algoritmo di calcolo molto efficiente. Inoltre, tutti i calcoli richiesti sono calcoli a farfalla che corrispondono essenzialmente a DFT su due punti. Per queste ragioni gli algoritmi validi nel caso di N potenza di 2 sono particolarmente semplici da mettere in pratica, e spesso nelle applicazioni è vantaggioso aver sempre a che fare con sequenze la cui durata è una potenza di 2. Ciò può essere fatto in molti casi semplicemente aumentando, se necessario, una sequenza di durata finita con valori nulli. Tuttavia, in alcuni

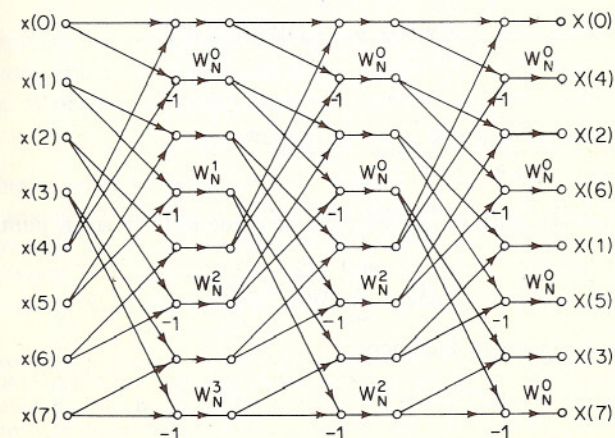


Fig. 6.23 Riordinamento della fig. 6.18 caratterizzato dalla stessa geometria per ogni stadio, per cui è possibile accedere ai dati e memorizzarli sequenzialmente. (Trasposta della fig. 6.14).

casi può non essere possibile scegliere un N che sia una potenza di 2, da cui la necessità di considerare la situazione più generale rappresentata dalla (6.25). Consideriamo perciò l'applicazione del principio della decimazione nel tempo al caso in cui N sia un prodotto di fattori non tutti necessariamente uguali a 2.

Definiamo

$$q_1 = p_2 p_3 \dots p_v$$

così che

$$N = p_1 \cdot q_1$$

Se N è una potenza di 2, potremmo scegliere $p_1 = 2$ e $q_1 = N/2$. Usando la decimazione nel tempo dovremmo quindi scomporre la sequenza $x(n)$ in due sequenze, ciascuna contenente $(N/2)$ campioni, consistenti, rispettivamente, dei campioni di posto pari e dispari. Quando $N = p_1 \cdot q_1$, possiamo dividere la sequenza di ingresso in p_1 sequenze di q_1 campioni ciascuna associando ogni p_1 -mo campione con una data sottosequenza. Ad esempio, se $p_1 = 3$ e $q_1 = 4$, così che $N = 12$, possiamo scomporre $x(n)$ in tre sequenze di lunghezza 4, con la prima sequenza consistente dei campioni $x(0)$, $x(3)$, $x(6)$, $x(9)$; la seconda sequenza consistente di $x(1)$, $x(4)$, $x(7)$, $x(10)$; e la terza sequenza consistente di $x(2)$, $x(5)$, $x(8)$ e $x(11)$. In generale possiamo scrivere $X(k)$ come

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n) W_N^{kn} \\ &= \sum_{r=0}^{q_1-1} x(p_1 r) W_N^{p_1 r k} + \sum_{r=0}^{q_1-1} x(p_1 r + 1) W_N^{(p_1 r + 1)k} + \dots \\ &\quad + \sum_{r=0}^{q_1-1} x(p_1 r + p_1 - 1) W_N^{(p_1 r + p_1 - 1)k} W_N^{p_1 r k} \end{aligned}$$

o

$$X(k) = \sum_{l=0}^{p_1-1} W_N^{lk} \sum_{r=0}^{q_1-1} x(p_1 r + l) W_N^{p_1 r k} \quad (6.26)$$

Le somme interne possono essere espresse come le DFT su q_1 punti

$$G_l(k) = \sum_{r=0}^{q_1-1} x(p_1 r + l) W_{q_1}^{rk} \quad (6.27)$$

poiché, come si verifica facilmente, è

$$W_N^{p_1 r k} = W_{q_1}^{rk} \quad \text{per} \quad N = p_1 \cdot q_1 \quad (6.28)$$

Perciò la (6.26) esprime $X(k)$ in termini di p_1 trasformate di Fourier discrete di sequenze lunghe q_1 campioni. Per determinare il numero di moltiplicazioni ed addizioni complesse richieste per effettuare la DFT secondo

la (6.26), consideriamo, come abbiamo fatto nella discussione originaria della decimazione nel tempo, il caso in cui le DFT su q_1 punti siano effettuate per mezzo del calcolo diretto. Dalla (6.26) osserviamo che il numero di DFT da calcolare su q_1 punti è p_1 . Perciò sono richieste un numero totale di $p_1 \cdot q_1^2$ moltiplicazioni e somme complesse. La somma esterna nella (6.26) è effettuata moltiplicando le DFT su q_1 punti per il fattore W_N^{lk} e sommando insieme i risultati. Poiché la sommatoria doppia nella (6.26) deve essere effettuata per N valori di k , sono necessarie un totale di $N(p_1 - 1)$ moltiplicazioni e somme complesse per combinare le p_1 DFT su q_1 punti.⁵ Perciò, il numero totale di moltiplicazioni e somme complesse richiesto per calcolare la trasformata di Fourier discreta nella forma della (6.26) è $N(p_1 - 1) + p_1 q_1^2$. Osserviamo ora che le DFT su q_1 punti possono essere scomposte in modo analogo. In particolare, se ora rappresentiamo q_1 come

$$q_1 = p_2 \cdot q_2$$

le sequenze di q_1 punti nella somma interna della (6.26) possono essere spezzate in p_2 sottosequenze, ciascuna di q_2 punti, in modo che la somma interna nella (6.26) possa essere sostituita da una doppia sommatoria, analogamente a quanto fatto all'inizio. Quando si sia fatto questo, il numero di operazioni richiesto per calcolare le DFT su q_1 punti nella (6.26) è, invece di q_1^2 ,

$$q_1(p_2 - 1) + p_2 q_2^2 \quad (6.29)$$

Di conseguenza, il fattore q_1^2 nell'espressione $N(p_1 - 1) + p_1 q_1^2$ è sostituito dalla (6.29), e perciò il numero totale di moltiplicazioni e somme complesse richieste è

$$N(p_1 - 1) + N(p_2 - 1) + p_1 p_2 q_2^2 \quad (6.30)$$

Se continuiamo questo procedimento scomponendo ulteriormente le DFT su q_2 punti, allora, quando la sequenza originaria è stata scomposta il più possibile, il numero di moltiplicazioni e somme complesse sarà

$$N(p_1 + p_2 + \dots + p_v - v) \quad (6.31)$$

Ad esempio, quando $p_1 = p_2 = \dots = p_v = p$, il numero di moltiplicazioni e somme complesse è $N(p - 1)v$. Quando $p = 2$, questo numero è $N \cdot v$, come già noto⁶. In generale, si può vedere dalla (6.31) che è preferibile protrarre la scomposizione sulla base di quanti più fattori possibile per un dato N . Formalmente, non c'è alcun vantaggio nello scegliere

⁵ Sommare p_1 termini richiede $p_1 - 1$ addizioni e non c'è bisogno di moltiplicare per W_N^{lk} quando $l = 0$. Ricordiamo al lettore che in tutto questo capitolo abbiamo di solito contato la moltiplicazione per W_N^{lk} anche quando W_N^{lk} vale uno o j . Nell'interpretare la relazione (6.26), invece, è opportuno tenere presente che W_N^{lk} vale uno per $l = 0$, affinché il risultato che otteniamo qui sia in accordo con la discussione del par. 6.2.

⁶ Ricordiamo che il numero delle moltiplicazioni può essere ulteriormente diminuito sfruttando certe proprietà di simmetria.

re fattori non primi, poiché se $p_i = r_i \cdot s_i$, con r_i e $s_i > 1$, allora $p_i > r_i + s_i$, tranne quando $r_i = s_i = 2$, nel qual caso $p_i = r_i + s_i$. Tuttavia, vi sono esempi (segnatamente $p_i = 4$ o 8) in cui si hanno ulteriori risparmi che non sono contemplati dalla (6.31).

Per illustrare il procedimento della decimazione nel tempo con N non potenza di 2, consideriamo il calcolo di una DFT su 18 punti, cioè con $N = 3 \cdot 3 \cdot 2$. Ponendo $p_1 = 3$ e $q_1 = 6$, e seguendo il procedimento visto sopra, dividiamo dapprima la sequenza originaria in tre sequenze, ciascuna della lunghezza di sei punti:

Sequenza 1: $x(0) \ x(3) \ x(6) \ x(9) \ x(12) \ x(15)$

Sequenza 2: $x(1) \ x(4) \ x(7) \ x(10) \ x(13) \ x(16)$

Sequenza 3: $x(2) \ x(5) \ x(8) \ x(11) \ x(14) \ x(17)$

Dividendo la sequenza originaria in queste tre sottosequenze, possiamo esprimere $X(k)$ come

$$X(k) = \sum_{l=0}^2 W_{18}^{lk} \sum_{r=0}^5 x(3r + l) W_{18}^{3rk} \quad (6.32)$$

$$= G_0(k) + W_{18}^k G_1(k) + W_{18}^{2k} G_2(k)$$

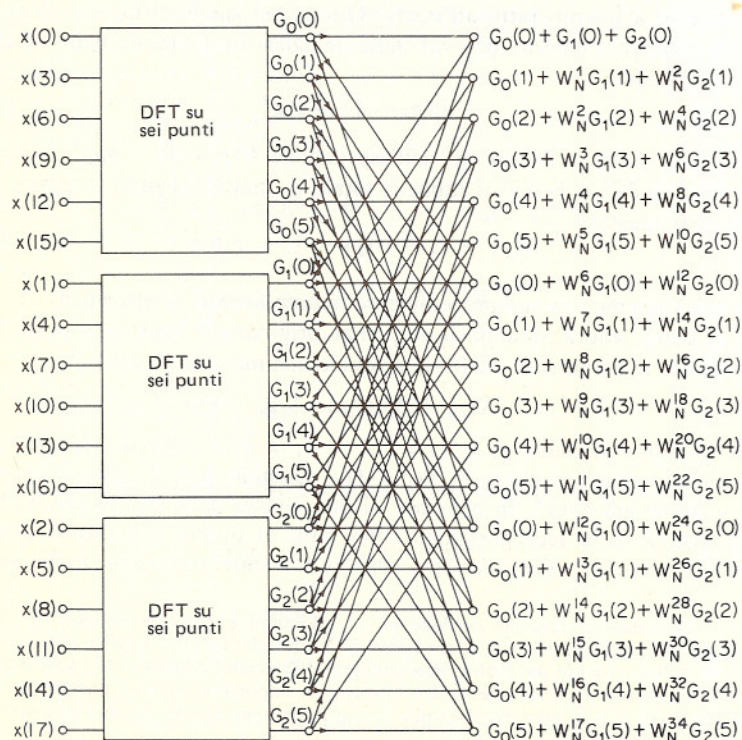


Fig. 6.24 Grafo di flusso del primo stadio della scomposizione di una DFT su 18 punti.

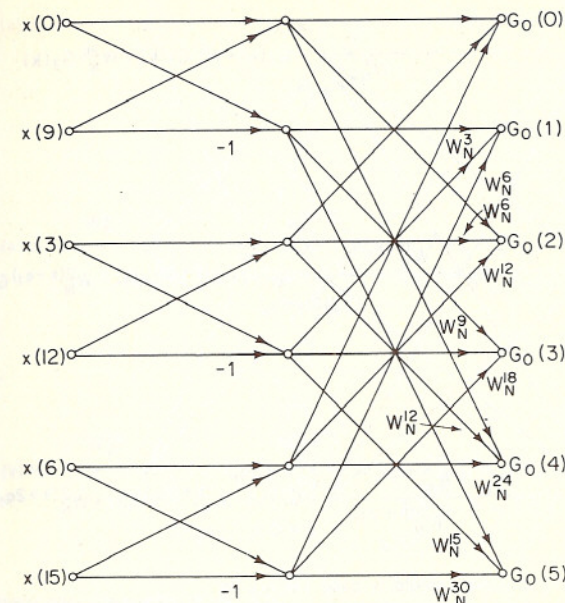


Fig. 6.25 Grafo di flusso dell'ulteriore scomposizione di una delle DFT su sei punti di fig. 6.24.

La somma più interna è una DFT su sei punti, con $l = 0$ corrispondente alla sequenza 1, $l = 1$ corrispondente alla sequenza 2, ed $l = 2$ corrispondente alla sequenza 3. In questo caso le DFT su sei punti $G_l(k)$ sono periodiche con periodo 6. Il calcolo dell'espressione (6.32) è illustrato in fig. 6.24.⁷

Le DFT su sei punti corrispondenti alla somma più interna nella (6.32) possono essere ulteriormente scomposte spezzando le sequenze $x(3r + l)$ in tre sequenze, ciascuna di due punti, o, in alternativa, in due sequenze, ciascuna di tre punti. Scegliendo il primo modo, cioè spezzando ciascuna delle sottosequenze di sei punti in tre sequenze di due punti, sostituiamo la somma più interna con

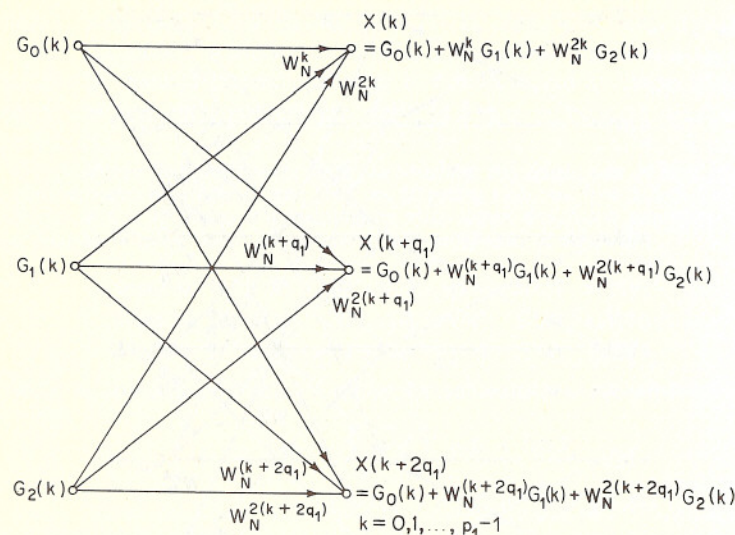
$$G_l(k) = \sum_{r=0}^5 x(3r + l) W_6^{rk} = \sum_{s=0}^2 W_6^{sk} \sum_{p=0}^1 x(9p + 3s + l) W_6^{3pk}$$

così che il calcolo complessivo di $X(k)$ diviene

$$X(k) = \sum_{l=0}^2 W_{18}^{lk} \sum_{s=0}^2 W_6^{sk} \sum_{p=0}^1 x(9p + 3s + l) W_6^{3pk} \quad (6.33)$$

Una delle DFT su sei punti ($G_0(k)$) è mostrata in dettaglio in fig. 6.25. Le altre due hanno forma identica. Quando la fig. 6.25 e i grafi di flusso corrispondenti a $G_1(k)$ e $G_2(k)$ sono posti nelle posizioni appropriate in fig. 6.24, la sequenza di ingresso è nell'ordine: $x(0), x(9), x(3), x(12), x(6), x(15), x(1), x(10), x(4), x(13), x(7), x(16), x(2), x(11), x(5), x(14), x(8)$ e $x(17)$. Osserviamo dalle fig. 6.24 e 6.25 che con questo ordinamento degli ingressi, il calcolo può essere fatto sul posto. Le trasformate su due punti sono indicate dalle famigliari farfalle nel primo stadio di fig. 6.25; l'operazione base della DFT su tre punti è un po' più complicata, ma è ancora, ovviamente,

⁷ In questa figura i coefficienti di trasmissione di ogni ramo devono essere dedotti dall'espressione algebrica associata ad ogni nodo d'uscita.

Fig. 6.26 Grafo di flusso del calcolo di base per $N=3q_1$.

un calcolo sul posto. Invece che a disposizione invertita dei bit, l'ordinamento dell'ingresso è un po' più complicato: precisamente, se indichiamo con $X_0(\cdot)$ la sequenza di ingresso, si può mostrare che

$$X_0(6l + 2s + p) = x(9p + 3s + l)$$

dove $p = 0, 1, 2$; $s = 0, 1, 2$; ed $l = 0, 1, 2$. Cioè, per effettuare il calcolo sul posto, l'ingresso deve essere memorizzato in un ordine « a disposizione invertita delle cifre » generalizzato. Come è evidente dalla fig. 6.24, l'uscita risultante è in ordine normale. Notiamo dalla fig. 6.24 che il calcolo base nell'ultimo stadio (come per fattori di 3 in generale) è quello illustrato in fig. 6.26, valido per $N = 3 \cdot q_1$.

Ricordiamo che nel caso di $N = 2q_1$ sapevamo dimezzare il numero di moltiplicazioni sfruttando certe proprietà di simmetria. Nel caso di $N = 3q_1$, come in fig. 6.26, una analoga manipolazione del grafo di flusso porta alla fig. 6.27. Poiché $N = 3q_1$, il moltiplicatore complesso di base $W_N^{q_1}$ è

$$W_N^{q_1} = e^{-j(2\pi/3)}$$

Perciò, $W_N^{q_1}$ e tutte le sue potenze sono coefficienti complessi che richiedono moltiplicazioni. Quindi la fig. 6.27 non è più efficiente della fig. 6.26.

Se invece di $N = 3q_1$ si considera il caso $N = 4q_1$, si può mostrare (v. il probl. 7 di questo capitolo) che il calcolo base della DFT diventa quello illustrato dal grafo di flusso di fig. 6.28. Tale grafo può essere ridisegnato come in fig. 6.29, con un conseguente risparmio di almeno 9 moltiplicazioni complesse sulle 12 mostrate in fig. 6.28. Analoghi risparmi si hanno per fattori 8, 16, ecc. [11]. Perciò, anche se $N = 2^v$, è a volte vantaggioso basare il calcolo su fattori 4, usando uno stadio basato sul fattore 2 se v è dispari.

La discussione svolta in questo paragrafo, sebbene parallela a quella precedente, è stata tutt'altro che completa, nel senso che abbiamo soltanto

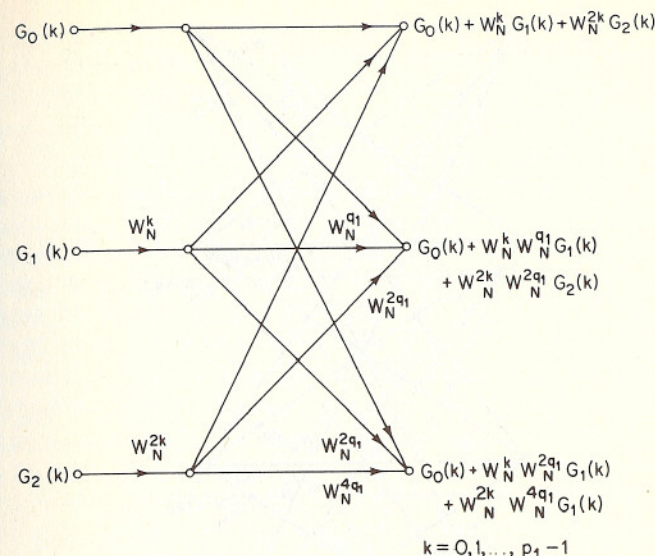


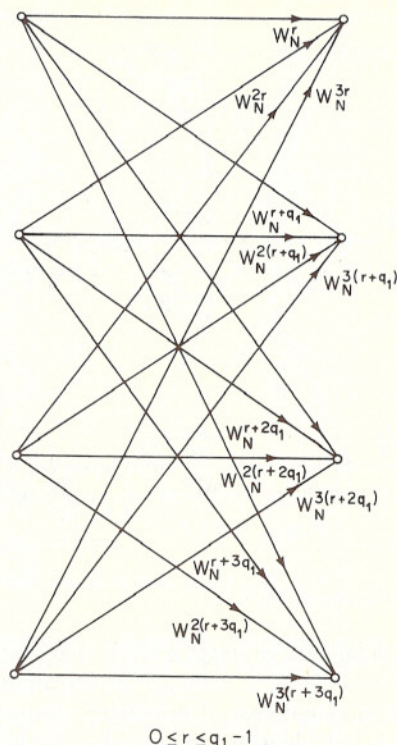
Fig. 6.27 Disposizione alternativa di fig. 6.26.

tentato di indicare alcuni dei vantaggi e degli svantaggi dell'uso di valori di N con fattori diversi da 2. I vantaggi principali sono gli aumenti di flessibilità e velocità in alcuni casi; il principale svantaggio è il grande aumento di complessità dell'algoritmo di calcolo. Anche se abbiamo preso in esame solo la decimazione nel tempo, discorsi analoghi si possono fare per la decimazione in frequenza. Per una discussione più dettagliata di algoritmi FFT nel caso di N numero composto qualsiasi, si veda Gentleman e Sande [9] e Singleton [10].

6.5 CONSIDERAZIONI GENERALI SU PROBLEMI DI CALCOLO PER GLI ALGORITMI DI FFT

Abbiamo discusso i principi base per il calcolo efficiente delle trasformate di Fourier discrete. In questa discussione abbiamo preferito usare le rappresentazioni in termini di grafi di flusso di segnale piuttosto che la scrittura dettagliata delle equazioni che tali grafi di flusso rappresentano. Per necessità abbiamo preso in esame grafi di flusso per valori particolari di N . Questo approccio si giustifica in base al fatto che tali grafi di flusso sono facilmente generalizzabili a valori di N qualsiasi. Ad esempio, considerando un grafo di flusso come quello di fig. 6.10, è possibile capire come strutturare un algoritmo generale di calcolo applicabile a qualsiasi $N = 2^v$.

Dall'esame dei grafi di flusso dei paragrafi precedenti risulta evidente che in ogni algoritmo di FFT esistono due aspetti principali. Il primo riguarda l'accesso e la memorizzazione dei dati negli stadi intermedi. Il

Fig. 6.28 Grafo di flusso del calcolo di base per $N=4q_1$.

secondo riguarda l'effettiva realizzazione del calcolo della farfalla una volta ottenuti i dati necessari. Sebbene i grafi di flusso dei paragrafi precedenti mettano in evidenza le operazioni essenziali dei rispettivi algoritmi di FFT, vi sono molti dettagli collegati ai due aspetti suddetti che occorre considerare nella realizzazione di un algoritmo di FFT. In questo paragrafo ci proponiamo di discutere alcuni di tali dettagli. Come già fatto in precedenza, concentreremo la nostra attenzione sugli algoritmi per $N=2^n$, sebbene buona parte della discussione valga anche nel caso più generale.

6.5.1 Indici

Consideriamo a titolo di esempio l'algoritmo illustrato in fig. 6.10. In questo caso l'ingresso deve essere in ordine a disposizione invertita dei bit in modo che il calcolo possa essere effettuato sul posto. Il risultato è allora in ordine normale. In generale, le sequenze non si presentano in ordine a disposizione invertita dei bit, così che il primo passo nella realizzazione di fig. 6.10 dovrà essere un processo di permutazione dei campioni della se-

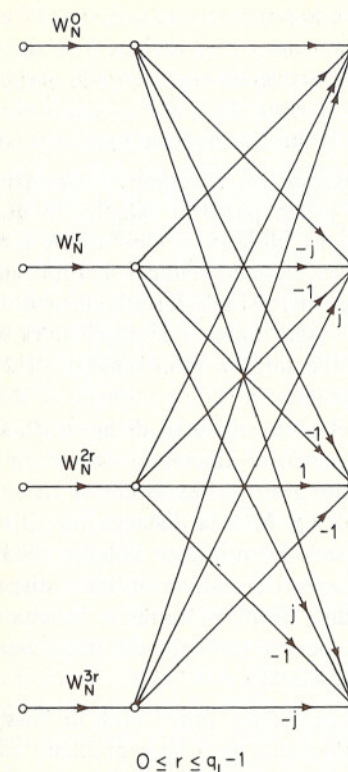


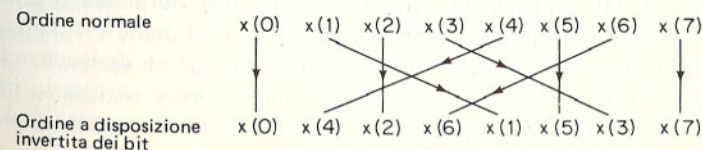
Fig. 6.29 Disposizione alternativa di fig. 6.28, risultante in un risparmio di moltiplicazioni.

quenza di ingresso. Questo processo è illustrato in fig. 6.30 per $N=8$.

Dalla figura risulta chiaro che il riordinamento di $x(n)$ a disposizione invertita dei bit può essere effettuato sul posto. Ciò è fattibile usando un indice che « conti » in ordine a disposizione invertita dei bit (un grafo di flusso di tale contatore a disposizione invertita dei bit è dato da Gold e Rader [7]). Si supponga che n sia l'indice normale e che l sia l'indice a disposizione invertita dei bit. Allora la sequenza d'ingresso è

$$X_0(l) = x(n)$$

Osserviamo dalla fig. 6.30 che per $n=l$ non è necessario alcuno scambio, ma quando $n \neq l$ dobbiamo scambiare $x(n)$ e $x(l)$. Dobbiamo però assicurarci che lo scambio sia fatto una sola volta e a questo fine si può confron-

Fig. 6.30 Riordinamento a disposizione invertita dei bit per $N=8$.

tare n con l ed effettuare lo scambio solo quando $l > n$. Perciò un semplice algoritmo per il riordinamento a disposizione invertita dei bit è il seguente. Apporre degli indici alla sequenza $x(n)$, facendo variare n da 0 a $N - 1$ in ordine normale, mentre l varia da 0 a $N - 1$ in ordine a disposizione invertita dei bit. Ogniquale volta $l > n$, scambiare $x(n)$ ed $x(l)$.

Una volta che l'ingresso sia in ordine a disposizione invertita dei bit, possiamo procedere al primo passo di calcolo. In questo caso i moltiplicatori sono tutti $W_N^0 = 1$ e gli ingressi alle farfalle sono elementi adiacenti della sequenza $X_0(\cdot)$. Nel secondo stadio i moltiplicatori sono tutti o W_N^0 o potenze di $W_N^{N/4}$, e gli ingressi alle farfalle distano di due posizioni nella sequenza $X_1(\cdot)$. Nello stadio m -mo i moltiplicatori sono una potenza di $W_N^{(N/2^m)}$ e gli ingressi alle farfalle sono separati di 2^{m-1} . Notiamo che le potenze di W_N sono richieste in ordine normale se il calcolo delle farfalle comincia dalla cima del grafo di flusso di fig. 6.10. Quanto enunciato in precedenza definisce il modo di accesso ai dati in un certo stadio. Questo dipende naturalmente dal grafo di flusso che si sta realizzando. Ad esempio, allo stadio m -mo di fig. 6.12 la distanza tra gli ingressi delle farfalle è $2^{(v-m)}$, i coefficienti sono nuovamente potenze di $W_N^{(N/2^m)}$, e in questo caso le potenze di W_N sono richieste in ordine a disposizione invertita dei bit. L'ingresso è in ordine normale; tuttavia, l'uscita è in ordine a disposizione invertita dei bit, per cui potrebbe essere necessario riordinare l'uscita come discusso in precedenza.

In generale, se consideriamo tutti i grafi di flusso dei par. 6.2 e 6.3, vediamo che ciascun algoritmo ha i suoi problemi caratteristici per l'uso degli indici. La scelta di un particolare algoritmo dipende da numerosi fattori. Gli algoritmi che utilizzano un calcolo sul posto hanno il vantaggio di sfruttare in maniera efficiente la memoria; d'altra parte, il tipo di memoria richiesta è ad accesso casuale invece che sequenziale. Questi algoritmi hanno in più lo svantaggio che o i dati in ingresso o i punti in uscita sono in ordine a disposizione invertita dei bit. Inoltre, a seconda che si scelga un algoritmo a decimazione nel tempo o a decimazione in frequenza e che gli ingressi o le uscite siano in ordine a disposizione invertita dei bit, i coefficienti devono essere presi in ordine normale o in ordine a disposizione invertita dei bit. Se non si ha a disposizione una memoria ad accesso casuale, ma si dispone invece di memoria ad accesso sequenziale, abbiamo già visto che esistono algoritmi di FFT che utilizzano questo tipo di memoria, ma ancora o l'ingresso o l'uscita devono essere in ordine a disposizione invertita dei bit. Anche se è possibile riordinare il grafo di flusso dell'algoritmo in modo tale che gli ingressi, le uscite e i coefficienti siano tutti in ordine normale, tuttavia i problemi che ne derivano relativi alla gestione degli indici sono alquanto complessi ed è necessario impiegare il doppio di memoria ad accesso casuale. Di conseguenza, l'uso di questi algoritmi non sembra vantaggioso.

Gli algoritmi di FFT di uso più comune sono quelli delle fig. 6.10, 6.12, 6.18 e 6.21, per i quali il calcolo è sul posto. Se i dati devono essere trasformati solo una volta, allora chiaramente occorre effettuare l'ordinamento a disposizione invertita dei bit o all'ingresso o all'uscita. Tuttavia, in alcune situazioni, è necessario trasformare i dati, modificare il risultato in qualche modo, e quindi antitrasformare. Ad esempio, nella realizzazione di filtri numerici FIR tramite la trasformata di Fourier discreta, si calcola la trasformata di un segmento dei dati di ingresso, la si moltiplica per la DFT della risposta all'impulso del filtro e si antitrasforma il risultato. Come altro esempio, nel calcolo di una funzione di autocorrelazione o di correlazione incrociata per mezzo della trasformata di Fourier discreta, si trasformeranno i dati, si moltiplicheranno le DFT, e quindi si antitrasformerà il prodotto risultante.

Quando due trasformate sono messe in cascata in tal modo, è possibile evitare la necessità dell'inversione dei bit con una scelta appropriata degli algoritmi di FFT. Ad esempio, nella realizzazione di un filtro numerico FIR per mezzo della DFT, possiamo scegliere un algoritmo per la trasformata diretta che utilizzi i dati in ordine normale e fornisca una DFT in ordine a disposizione invertita dei bit. Sia il grafo di flusso della fig. 6.12, basato sulla decimazione nel tempo, che quello di fig. 6.18, basato sulla decimazione in frequenza, potrebbero essere usati in questo modo. La differenza tra queste due forme è che quella a decimazione nel tempo richiede i coefficienti in ordine a disposizione invertita dei bit, mentre la forma a decimazione in frequenza richiede i coefficienti in ordine normale.

Usando l'uno o l'altro di tali algoritmi, la trasformata si ottiene in ordine a disposizione invertita dei bit e, di conseguenza, avremo memorizzato la DFT corrispondente alla risposta in frequenza del filtro in ordine a disposizione invertita dei bit. Per la DFT inversa possiamo quindi scegliere una forma dell'algoritmo che abbia dati in ingresso in ordine a disposizione invertita dei bit e fornisca risultati in ordine normale. Qui possono essere usati sia il grafo di flusso di fig. 6.10, basato sulla decimazione nel tempo, che quello di fig. 6.21, basato sulla decimazione in frequenza. La fig. 6.10, tuttavia, utilizza i coefficienti in ordine normale, mentre la fig. 6.21 richiede i coefficienti in ordine a disposizione invertita dei bit. Per far sì che i coefficienti debbano essere forniti solo o in ordine normale (che è preferibile), o in ordine a disposizione invertita dei bit ci sono due possibilità. Se si sceglie l'algoritmo della decimazione nel tempo per la trasformata diretta, allora si dovrebbe scegliere l'algoritmo della decimazione in frequenza per la trasformata inversa, col che i coefficienti devono essere in ordine a disposizione invertita dei bit. In alternativa, si può accoppiare l'algoritmo della decimazione in frequenza per la trasformata diretta con l'algoritmo della decimazione nel tempo per la trasformata inversa, utilizzando quindi i coefficienti in ordine normale.

6.5.2 Coefficienti

Abbiamo osservato che i coefficienti W_N^r possono essere richiesti in ordine a disposizione invertita dei bit o in ordine normale. In entrambi i casi, dobbiamo memorizzare una tabella sufficientemente estesa in cui cercare tutti i valori richiesti oppure calcolare i valori che servono. La prima alternativa ha il vantaggio della velocità, ma naturalmente richiede più memoria. Osserviamo dai grafi di flusso che abbiamo bisogno di W_N^r per $r = 0, 1, \dots, N/2 - 1$. Perciò abbiamo bisogno di $(N/2)$ posizioni di memoria complesse per una tabella completa di valori di W_N^r ⁸. Nel caso di algoritmi che richiedono i coefficienti in ordine a disposizione invertita dei bit, è necessario formare la tabella di conseguenza.

Il calcolo dei coefficienti man mano che servono permette un risparmio di memoria, ma è meno efficiente. L'efficienza maggiore si ottiene usando una formula ricorsiva. Notiamo che, in generale, ad un dato stadio i coefficienti sono tutti potenze di qualche potenza di W_N ; cioè, W_N^q , dove q dipende dall'algoritmo e dallo stadio. Perciò, se i coefficienti servono in ordine normale, possiamo usare la formula ricorsiva

$$W_N^{ql} = W_N^q \cdot W_N^{q(l-1)} \quad (6.34)$$

per ottenere il coefficiente l -esimo dall' $(l-1)$.mo. Chiaramente, gli algoritmi che richiedono i coefficienti in ordine a disposizione invertita dei bit non sono adatti per questo approccio. Si dovrebbe notare che la formula (6.34) è essenzialmente l'oscillatore in forma accoppiata del probl. 3 di questo capitolo. Quando si usa un'aritmetica a precisione finita, gli errori possono sommarsi nell'iterazione di questa equazione alle differenze. Perciò, è in generale necessario ri-inizializzare i valori in certi punti (ad es., $W_N^{N/4} = -j$) in modo che gli errori non diventino inaccettabili.

6.5.3 Trasformate di Fourier veloci multidimensionali

La trasformata di Fourier discreta bidimensionale è stata definita nel cap. 3 come

$$X(k, l) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m, n) W_M^{km} W_N^{ln} \quad (6.35)$$

$$k = 0, 1, \dots, M-1, l = 0, 1, \dots, N-1$$

Osserviamo che la (6.35) è molto simile alla (6.26). In effetti, è possibile interpretare un algoritmo di FFT in termini di DFT multidimensionali (v. Gold e Rader [7]). Se siamo interessati al calcolo della (6.35), possiamo osservare che esso comporta M DFT della forma

$$A(m, l) = \sum_{n=0}^{N-1} x(m, n) W_N^{ln}, \quad \begin{matrix} m = 0, 1, \dots, M-1 \\ l = 0, 1, \dots, N-1 \end{matrix} \quad (6.36)$$

⁸ Questo numero può essere ridotto sfruttando certe proprietà di simmetria, al prezzo di una maggiore complessità della ricerca dei valori desiderati.

seguite da N DFT della forma

$$X(k, l) = \sum_{m=0}^{M-1} A(m, l) W_M^{km}, \quad \begin{matrix} k = 0, 1, \dots, M-1 \\ l = 0, 1, \dots, N-1 \end{matrix} \quad (6.37)$$

Chiaramente, possiamo calcolare le (6.36) e (6.37) per mezzo di uno qualsiasi degli algoritmi prima descritti. Se M ed N sono potenze di 2, il numero di moltiplicazioni complesse necessario sarà

$$M \cdot \frac{N}{2} \log_2 N + N \cdot \frac{M}{2} \log_2 M = \frac{NM}{2} (\log_2 N \cdot M)$$

di fronte a circa $N^2 \cdot M^2$ moltiplicazioni complesse per il calcolo diretto della (6.35).

Una difficoltà fondamentale nei calcoli di DFT bidimensionali è l'ammontare di memoria necessaria per effettuare il calcolo sul posto. Per dati complessi, sono necessarie $2(N \cdot M)$ posizioni di memoria reali per memorizzare l'ingresso o la trasformata risultante. Ad esempio, se $N = M = 256$, sono necessarie 131,072 posizioni di memoria. Se non si ha a disposizione l'ammontare necessario di memoria ad accesso casuale, si può usare un qualunque tipo di memoria di massa ad accesso sequenziale, come disco o nastro, con le righe o le colonne memorizzate una dopo l'altra. Ciò aumenta la complessità di realizzazione della (6.36) o della (6.37). Ad esempio, se i dati sono memorizzati per righe, le trasformate di riga (6.36) possono essere effettuate facilmente, ma se queste sono memorizzate sequenzialmente per righe, risulta difficile accedere ai dati necessari per le trasformate di colonna (6.37). Un modo per ovviare a questa difficoltà consiste nell'effettuare parecchie trasformate di riga e memorizzare i valori risultanti di $A(m, l)$ in ordine trasposto. Dopo che si è formata tutta la tabella trasposta $A(l, m)$, le trasformate di colonna desiderate si ottengono calcolando trasformate di righe della tabella trasposta. La trasformata bidimensionale risultante va memorizzata in ordine trasposto [12].

6.6 ALGORITMO DELLA TRASFORMATTA Z CHIRP

Abbiamo visto come sia possibile calcolare la DFT in modo molto efficiente. Ciò corrisponde al calcolo efficiente dei campioni della trasformata z di una sequenza di lunghezza finita presi in punti equispaziati sulla circonferenza unitaria. Notiamo però che, per ottenere questa efficienza nel calcolo della trasformata z , N deve essere un numero composto di molti fattori. Inoltre, possiamo essere interessati al campionamento della trasformata z su un qualche altro percorso, o possiamo non volere campioni della trasformata z su tutta la circonferenza unitaria. Per questo motivo sono molto interessanti tutti quei metodi che accrescono la flessibilità degli algoritmi di calcolo (efficiente) della DFT, o ne estendono il campo di applicazione.

Supponiamo di voler ottenere campioni della trasformata z di una sequenza di durata finita su una circonferenza concentrica con la circonferenza unitaria, e che i campioni debbano essere equispaziati in angolo lungo questa circonferenza. In tal caso, con una minima modifica, si può usare un algoritmo di FFT: infatti, se abbiamo una sequenza di durata finita $x(n)$ di lunghezza N , la trasformata di Fourier discreta della sequenza $x(n)\alpha^{-n}$ fornirà N valori equispaziati lungo una circonferenza di raggio α nel piano z . Se vogliamo ottenere campioni in frequenza equispaziati su una piccola porzione della circonferenza unitaria, l'approccio più efficiente può spesso consistere nell'uso di un algoritmo di FFT per calcolare campioni in frequenza con la spaziatura desiderata, ottenendo però campioni al di fuori dell'intervallo di frequenza che interessa. Ad esempio, se avessimo una sequenza di 128 punti e fossimo interessati ad ottenere 128 campioni della trasformata z sulla circonferenza unitaria tra $\omega = -\pi/8$ e $\omega = +\pi/8$, la procedura più efficiente potrebbe essere calcolare una DFT su 1024 punti aumentando la sequenza originaria con zeri e tenere soltanto i 128 punti desiderati della trasformata.

Una procedura alternativa, che in molte situazioni è la più efficiente, è l'uso dell'algoritmo della trasformata z chirp (CZT) [13]. Questo algoritmo è orientato al calcolo di valori della trasformata z su di un percorso a spirale ed equispaziati in angolo su una qualche porzione della spirale. Specificamente, sia $x(n)$ una sequenza di N punti ed $X(z)$ la sua trasformata z . Usando l'algoritmo della CZT, $X(z)$ può essere calcolata nei punti z_k dati da

$$z_k = AW^{-k}, \quad k = 0, 1, \dots, M-1 \quad (6.38)$$

dove

$$W = W_0 e^{-j\phi_0}$$

$$A = A_0 e^{j\theta_0}$$

con W_0 ed A_0 numeri reali positivi. Di conseguenza, il percorso lungo il quale si ottengono i campioni è quello indicato in fig. 6.31.

Questo percorso è una spirale nel piano z . Il parametro W_0 controlla le caratteristiche della spirale: se W_0 è maggiore dell'unità, la spirale evolve verso l'origine al crescere di k , mentre se W_0 è inferiore all'unità, evolve verso l'esterno al crescere di k . I parametri A_0 e θ_0 sono la posizione in raggio ed angolo, rispettivamente, del primo valore, cioè, per $k = 0$. I restanti valori si trovano lungo il percorso della spirale con una spaziatura angolare di ϕ_0 . Di conseguenza, se $W_0 = 1$, la spirale è, in effetti, un arco di circonferenza, e se $A_0 = 1$, questo arco di circonferenza fa parte della circonferenza unitaria.

Con i valori di z_k dati dalla (6.38) vogliamo calcolare

$$X(z_k) = \sum_{n=0}^{N-1} x(n) A^{-n} W^{nk}, \quad k = 0, 1, \dots, M-1 \quad (6.39)$$

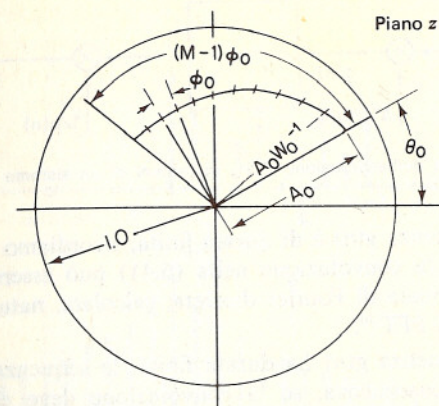


Fig. 6.31 Percorso nel piano z per la trasformata z chirp.

dove N è la lunghezza della sequenza $x(n)$. Usando l'identità⁹

$$nk = \frac{1}{2}[n^2 + k^2 - (k-n)^2] \quad (6.40)$$

la (6.39) può essere scritta come

$$X(z_k) = \sum_{n=0}^{N-1} x(n) A^{-n} W^{n^2/2} W^{k^2/2} W^{-(k-n)^2/2}$$

o

$$X(z_k) = W^{k^2/2} \sum_{n=0}^{N-1} x(n) A^{-n} W^{n^2/2} W^{-(k-n)^2/2}$$

Ponendo

$$g(n) = x(n) A^{-n} W^{n^2/2}$$

possiamo allora scrivere

$$X(z_k) = W^{k^2/2} \sum_{n=0}^{N-1} g(n) W^{-(k-n)^2/2}, \quad k = 0, 1, \dots, M-1 \quad (6.41)$$

Con $X(z_k)$ espressa nella forma (6.41), riconosciamo che la sommatoria corrisponde alla convoluzione della sequenza $g(n)$ con la sequenza $W^{-n^2/2}$. Perciò il calcolo dell'espressione (6.41) è come illustrato in fig. 6.32, dove

$$h(n) = W^{-n^2/2}$$

Quando A e W_0 sono unitari, la sequenza $h(n)$ può essere pensata come una sequenza esponenziale complessa con frequenza crescente linearmente. Nei sistemi radar, tali segnali vengono chiamati *segnali chirp*, da cui il nome *trasformata z chirp*. Un sistema simile alla fig. 6.32 è comunemente usato per l'analisi spettrale in problemi di radar.

⁹ Questo accorgimento è stato introdotto da Bluestein [14].

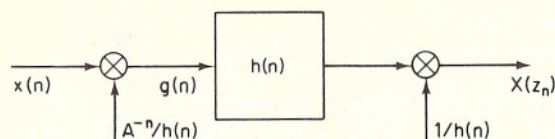


Fig. 6.32 Interpretazione dell'espressione (6.41) in termini di un sistema lineare.

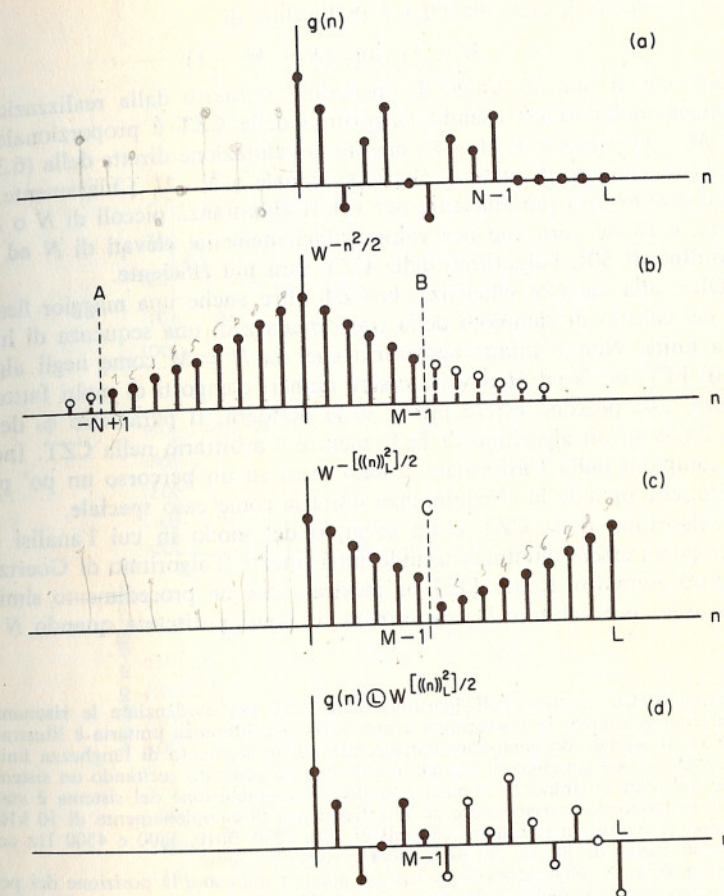
Poiché la sequenza $g(n)$ è di durata finita, ricordiamo dalla discussione del cap. 3 che la convoluzione nella (6.41) può essere effettuata per mezzo della trasformata di Fourier discreta, calcolata, naturalmente, usando un algoritmo di FFT¹⁰.

Mentre la sequenza $g(n)$ ha durata finita, la sequenza $W^{-n^2/2}$ ha durata infinita; di conseguenza, se la convoluzione deve essere realizzata usando la trasformata di Fourier discreta, è necessario segmentare la sequenza $W^{-n^2/2}$. Notiamo anche che, mentre il risultato della convoluzione ha lunghezza infinita, noi siamo invece interessati al risultato della convoluzione solo per $k = 0, 1, \dots, M-1$. Di conseguenza, segmentando la sequenza $W^{-n^2/2}$, sarebbe vantaggioso scegliere le sezioni in modo tale che il risultato del calcolo di una sezione fornisca gli M punti di uscita desiderati. La fig. 6.33 mostra le sequenze interessate in questo processo nel caso $N = 10$ ed $M = 6$. Le sequenze $g(n)$ e $W^{-n^2/2}$ sono illustrate, rispettivamente, in fig. 6.33(a) e (b).

Nella realizzazione della convoluzione di $g(n)$ con $W^{-n^2/2}$, la sola parte di $W^{-n^2/2}$ che è necessaria per calcolare il risultato della convoluzione nell'intervallo da 0 a $M-1$ è quella da $-N+1$ a $M-1$, estremi compresi. Tale parte della sequenza $W^{-n^2/2}$ è quella compresa tra le linee tratteggiate indicate con A e B in fig. 6.33(b). Di conseguenza, la convoluzione può essere realizzata calcolando la DFT su $(M+N-1)$ punti di $g(n)$ (aumentata naturalmente con $M-1$ zeri) e la DFT su $(M+N-1)$ punti della parte della sequenza $W^{-n^2/2}$ situata nella regione tra A e B in fig. 6.33(b). La trasformata inversa del prodotto di queste due trasformate di Fourier discrete sarà la convoluzione circolare della sequenza $g(n)$ con la porzione considerata di $W^{-n^2/2}$. Come già visto a proposito del metodo di convoluzione detto di sovrapposizione ed estrazione, una parte della convoluzione circolare corrisponderà ad una convoluzione lineare ed una parte no. Possiamo fare in modo che i punti « buoni » o desiderati cadano nella regione $0 \leq n \leq M-1$, interpretando l'indice n modulo $(N+M-1)$. Ciò significa che calcoleremo la DFT della sequenza

$$h(n) = \begin{cases} W^{-n^2/2}, & 0 \leq n \leq M-1 \\ W^{-(N+M-1-n)^2/2}, & M \leq n \leq N+M-2 \end{cases}$$

¹⁰ Bluestein [14] ha dimostrato che si può ottenere una realizzazione ricorsiva di fig. 6.32 per il caso $z_k = e^{j(2\pi/N)k}$ con N quadrato perfetto (v. probl. 16 di questo capitolo).

Fig. 6.33 Sequenze utilizzate nell'algoritmo della CZT ($L=N+M-1$).

come illustrato in fig. 6.33(c). Se moltiplichiamo le trasformate di Fourier discrete di $g(n)$ ed $h(n)$, i primi M valori della trasformata inversa corrispondente sono i valori desiderati della convoluzione di $g(n)$ con $W^{-n^2/2}$. Per ottenere gli M valori desiderati di $X(k)$ come nella (6.41), dobbiamo moltiplicare questi valori per $W^{-n^2/2}$.

In quanto si è visto finora la lunghezza delle DFT considerate era $(M+N-1)$. Se desideriamo calcolare la trasformata di Fourier discreta usando un algoritmo valido per lunghezze pari a potenze di 2, ciò può essere fatto facilmente aumentando le sequenze lunghe $(M+N-1)$ punti con un numero sufficiente di zeri in modo che la loro lunghezza totale sia una potenza di 2¹¹. Poiché il numero di moltiplicazioni complesse richie-

¹¹ Gli zeri devono essere inseriti nel punto C di fig. 6.33 (c).

ste per il calcolo di ciascuna DFT è dell'ordine di

$$(N + M - 1) \cdot \log_2 (N + M - 1)$$

è chiaro che il numero totale di operazioni richiesto dalla realizzazione del calcolo della (6.39) usando l'algoritmo della CZT è proporzionale a $(N + M - 1) \cdot \log_2 (N + M - 1)$, mentre la valutazione diretta della (6.39) richiede un numero di operazioni proporzionale a $N \cdot M$. Chiaramente, il metodo diretto sarà più efficiente per valori abbastanza piccoli di N o M ; tuttavia, è anche vero che per valori sufficientemente elevati di N ed M (dell'ordine di 50), l'algoritmo della CZT sarà più efficiente.

Oltre alla maggior efficienza, la CZT offre anche una maggior flessibilità nel calcolo di campioni della trasformata z di una sequenza di lunghezza finita. Non è infatti necessario che sia $N = M$ come negli algoritmi di FFT, né N ed M devono essere numeri composti di molti fattori; in effetti, essi possono essere primi, se si desidera. Il parametro ϕ_0 deve essere $2\pi/N$ in un algoritmo di FFT, mentre è arbitrario nella CZT. Inoltre, i campioni della trasformata z sono presi su un percorso un po' più generale, che include la circonferenza unitaria come caso speciale.

L'algoritmo della CZT è un esempio del modo in cui l'analisi di Fourier possa essere effettuata usando filtri lineari (l'algoritmo di Goertzel è un altro esempio). Rader [15] ha mostrato che un procedimento simile si può usare per valutare la trasformata di Fourier discreta quando N è primo.

ESEMPIO. Un esempio dell'algoritmo della CZT per evidenziare le risonanze di un sistema valutando la trasformata z non sulla circonferenza unitaria è illustrato in fig. 6.34. Il segnale da analizzare corrisponde ad un segmento di lunghezza finita di un segnale vocale sintetico. Il segnale vocale è stato generato eccitando un sistema a cinque poli con un treno di impulsi periodico. La simulazione del sistema è stata effettuata in modo da corrispondere ad una frequenza di campionamento di 10 kHz. I poli sono stati posti a frequenze centrali di 270, 2290, 3010, 3500 e 4500 Hz con larghezze di banda di 30, 50, 60, 87 e 140 Hz rispettivamente.

La fig. 6.34 (a) rappresenta il grafico nel piano z indicante la posizione dei poli usati per generare il segnale. L'algoritmo della CZT è stato applicato ad un periodo dei dati a regime per cinque diverse scelte di $|W|$ con i risultati mostrati in fig. 6.34 (b). I primi due spettri corrispondono a percorsi a spirale al di fuori della circonferenza unitaria con un risultante allargamento dei picchi di risonanza. $|W| = 1$ corrisponde alla valutazione della trasformata z sulla circonferenza unitaria. Al crescere di $|W|$ oltre l'unità, il percorso si svolge all'interno della circonferenza unitaria e più vicino alle posizioni dei poli col risultato di acuire i picchi di risonanza.

SOMMARIO

In questo capitolo abbiamo esaminato diverse tecniche per il calcolo della trasformata di Fourier discreta. Il nostro obiettivo è stato quello di mostrare come la periodicità e la simmetria del fattore complesso $e^{-j(2\pi/N)kn}$ possano essere sfruttate per aumentare l'efficienza dei calcoli della DFT.

Abbiamo considerato l'algoritmo di Goertzel e il calcolo diretto della DFT a causa dell'importanza di tali tecniche nel caso in cui non si vogliano tutti gli N valori della DFT.

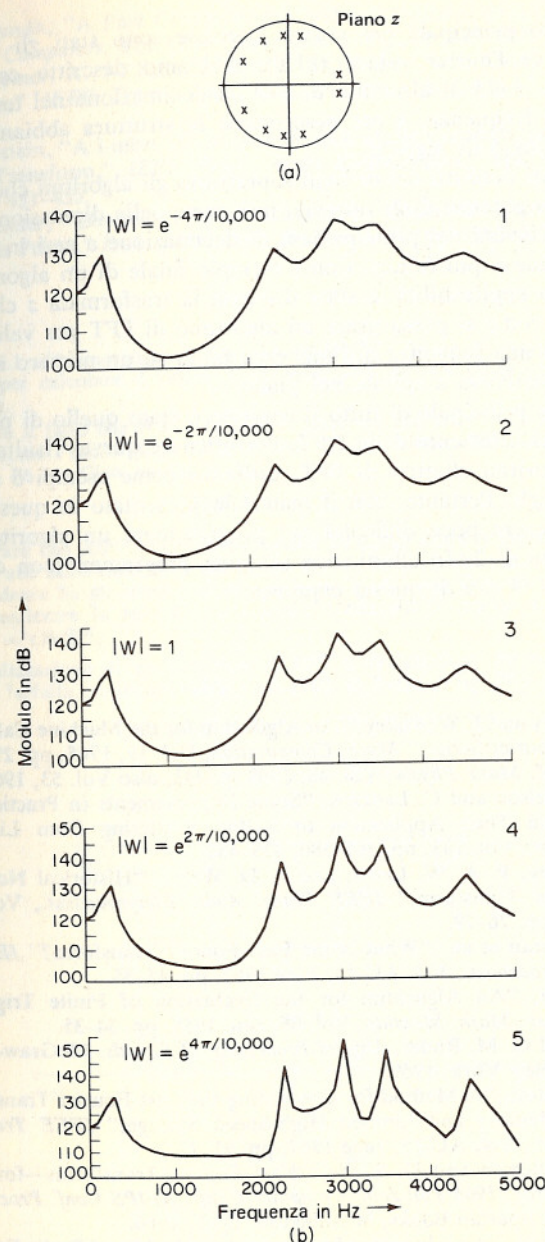


Fig. 6.34 Uso dell'algoritmo di CZT. (a) Posizioni dei poli nel piano z per un segnale vocale sintetico. (b) Valutazione della trasformata z per alcuni percorsi a spirale (da Rabiner, Schafer e Rader [13]).

L'oggetto principale del nostro interesse sono stati gli algoritmi di trasformata di Fourier veloce (FFT). Abbiamo descritto con un certo dettaglio due classi di algoritmi di FFT, a decimazione nel tempo e a decimazione in frequenza, e per descriverne la struttura abbiamo fatto uso dei grafi di flusso di segnale.

Sono stati descritti nei dettagli soprattutto gli algoritmi che richiedono che N sia una potenza di 2; tuttavia, una parte della discussione ha riguardato le applicazioni dei principi base di decimazione a casi in cui N sia il prodotto di due o più fattori. Come esempio finale di un algoritmo veloce con più vasta applicabilità, è stata discussa la trasformata z chirp. Abbiamo mostrato come si possa usare un algoritmo di FFT per valutare la trasformata z di una sequenza di lunghezza finita su un numero arbitrario di punti su un percorso a spirale nel piano z .

Lo scopo principale di tutto il capitolo è stato quello di presentare le basi del calcolo efficiente della DFT. Per ottenere questo risultato ci siamo serviti di algoritmi effettivi di FFT, utilizzati come esempi e descritti fin nei loro dettagli. Pertanto, con il materiale presentato in questo capitolo, si dovrebbe avere poca difficoltà nel programmare un algoritmo di FFT con N potenza di 2. In effetti, due semplici programmi (con errori) sono dati nei probl. 4 e 5 di questo capitolo.

BIBLIOGRAFIA

1. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Computation*, Vol. 19, 1965, pp. 297-301.
2. C. Runge, *Z. Math. Physik*, Vol. 48, 1903, p. 443; also Vol. 53, 1905, p. 117.
3. G. C. Danielson and C. Lanczos, "Some Improvements in Practical Fourier Analysis and Their Application to X-Ray Scattering from Liquids," *J. Franklin Inst.*, Vol. 233, pp. 365-380, 435-452.
4. J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "Historical Notes on the Fast Fourier Transform," *IEEE Trans. Audio Electroacoust.*, Vol. AU-15, June 1967, pp. 76-79.
5. W. T. Cochran et al., "What is the Fast Fourier Transform?" *IEEE Trans. Audio Electroacoust.*, Vol. AU-15, June 1967, pp. 45-55.
6. G. Goertzel, "An Algorithm for the Evaluation of Finite Trigonometric Series," *Amer. Math. Monthly*, Vol. 65, Jan. 1958, pp. 34-35.
7. B. Gold and C. M. Rader, *Digital Processing of Signals*. McGraw-Hill Book Company, New York, 1969.
8. R. C. Singleton, "A Method for Computing the Fast Fourier Transform with Auxiliary Memory and Limited High-Speed Storage," *IEEE Trans. Audio Electroacoust.*, Vol. AU-15, June 1967, pp. 91-97.
9. W. M. Gentleman and G. Sande, "Fast Fourier Transforms—for Fun and Profit," in *Proc. 1966 Fall Joint Computer Conf., AFIPS Conf. Proc.*, Vol. 29, pp. 563-578, Spartan Books, Washington, D.C., 1966.
10. R. C. Singleton, "An Algorithm for Computing the Mixed Radix Fast Fourier Transform," *IEEE Trans. Audio Electroacoust.*, Vol. AU-17, June 1969, pp. 93-103.
11. G. D. Bergland, "A Fast Fourier Transform Algorithm Using Base 8 Iterations," *Math. Computation*, Vol. 22, Apr. 1968, pp. 275-279.

12. J. O. Eklundh, "A Fast Computer Method for Matrix Transposing," *IEEE Trans. on Computers*, Vol. C-21, No. 7, July, 1972, pp. 801-803.
13. L. R. Rabiner, R. W. Schafer, and C. M. Rader, "The Chirp z -Transform Algorithm," *IEEE Trans. Audio Electroacoust.*, Vol. AU-17, June 1969, pp. 86-92.
14. L. I. Bluestein, "A Linear Filtering Approach to the Computation of Discrete Fourier Transform," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, Dec. 1970, pp. 451-455.
15. C. M. Rader, "Discrete Fourier Transforms When the Number of Data Samples is Prime," *Proc. IEEE*, Vol. 56, June 1968, pp. 1107-1108.

PROBLEMI

1. Nel par. 6.1 abbiamo usato il fatto che $W_N^{-kN} = 1$ per derivare un algoritmo ricorsivo per calcolare il valore $X(k)$ della DFT di una sequenza di lunghezza finita $x(n)$.

(a) Usando il fatto che $W_N^{kN} = W_N^{Nn} = 1$, mostrare che $X(N - k)$ può essere ottenuto come l'uscita dopo N iterazioni dell'equazione alle differenze illustrata in fig. P6.1-1. Cioè, mostrare che

$$X(N - k) = y_k(N)$$

(b) Mostrare che $X(N - k)$ è anche uguale all'uscita dopo N iterazioni dell'equazione alle differenze illustrata in fig. P6.1-2. Si dovrebbe notare che il sistema precedente ha gli stessi poli di quello di fig. 6.2, ma il coefficiente necessario per realizzare lo zero è il complesso coniugato di quello di fig. 6.2. Cioè, $W_N^{-k} = (W_N^k)^*$.

2. Nella realizzazione di un algoritmo di FFT a decimazione nel tempo, il calcolo di base a farfalla è quello illustrato nel grafo di flusso di fig. P6.2.

$$X_{m+1}(p) = X_m(p) + W_N^r X_m(q)$$

$$X_{m+1}(q) = X_m(p) - W_N^r X_m(q)$$

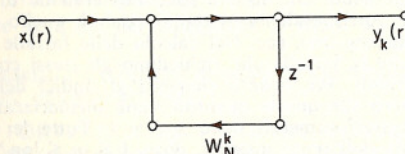


Fig. P6.1-1

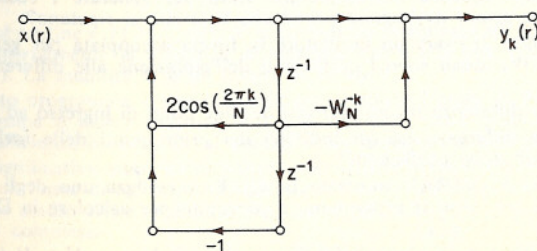


Fig. P6.1-2

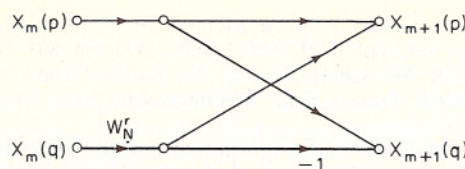


Fig. P6.2

Usando nella realizzazione dei calcoli un'aritmetica in virgola fissa, si suppone comunemente che tutti i numeri siano moltiplicati per un fattore di scala perché siano inferiori all'unità. Dobbiamo perciò esaminare il problema della saturazione (« overflow ») nei calcoli a farfalla.

(a) Mostrare che se si impone

$$|X_m(p)| < \frac{1}{2} \quad \text{e} \quad |X_m(q)| < \frac{1}{2}$$

non si può avere saturazione nei calcoli a farfalla; cioè,

$$|\operatorname{Re}[X_{m+1}(p)]| < 1, \quad |\operatorname{Im}[X_{m+1}(p)]| < 1, \quad |\operatorname{Re}[X_{m+1}(q)]| < 1,$$

e

$$|\operatorname{Im}[X_{m+1}(q)]| < 1.$$

(b) In pratica, è più facile e più conveniente imporre

$$|\operatorname{Re}[X_m(p)]| < \frac{1}{2}, \quad |\operatorname{Im}[X_m(p)]| < \frac{1}{2}$$

$$|\operatorname{Re}[X_m(q)]| < \frac{1}{2}, \quad |\operatorname{Im}[X_m(q)]| < \frac{1}{2}$$

Queste condizioni sono sufficienti ad assicurare che non si abbiano saturazioni nel calcolo della farfalla? Giustificare la risposta.

3. Nella realizzazione dell'algoritmo della FFT, è a volte utile generare le potenze di W_N ricorsivamente, usando un oscillatore in forma canonica o in forma accoppiata. Nella discussione che segue, considereremo l'algoritmo della FFT per N potenza di 2, realizzato nella forma caratterizzata dalla fig. 6.10, supponendo però in questo caso che N sia una potenza qualsiasi di 2.

Per generare efficientemente i coefficienti, la frequenza dell'oscillatore dovrebbe cambiare al cambiare dell'insieme di valori $X_m(l)$, cioè dello stadio di calcolo che si sta considerando. Gli insiemi sono numerati da 0 a $\log_2 N$, così che, ad esempio, l'insieme corrispondente ai dati iniziali è quello di ordine zero, il successivo quello di ordine uno, ecc. Nel calcolo delle farfalle all'interno di uno stadio, si valutano tutte le farfalle che richiedono gli stessi coefficienti prima di ottenere nuovi coefficienti. Per quanto riguarda gli indici delle quantità $X_m(l)$, per ogni m , si assumerà che queste quantità siano memorizzate in registri complessi (doppi) consecutivi, numerati da 0 a $N-1$. Tutte le seguenti domande sono riferite al calcolo dell' m -mo insieme, dove $1 \leq m \leq \log_2 N$. Le risposte dovrebbero essere in termini di m .

- Quante farfalle si devono calcolare?
 - Qual è la frequenza dell'oscillatore usato per generare i coefficienti, cioè, quante iterazioni avvengono prima che le uscite si ripetano?
 - Supponendo di usare un oscillatore in forma accoppiata per generare le potenze di W_N , quali sono i coefficienti dell'equazione alle differenze dell'oscillatore?
 - Qual è la differenza tra gli indirizzi dei due punti di ingresso ad una farfalla?
 - Qual è la differenza tra gli indirizzi dei primi punti delle farfalle che utilizzano gli stessi coefficienti?
4. Il programma FORTRAN mostrato in fig. P6.4 realizza uno degli algoritmi di FFT discussi nel testo. Il programma è concepito per calcolare la DFT,

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

```

SUBROUTINE FFT(X,M)
COMPLEX X(1024),U,W,T
N=2**M
NV2=N/2
NM1=N-1
J=1
DO 7 I=1,NM1
T=X(J)
X(J)=X(I)
X(I)=T
K=NV2
IF(K.GE.J) GO TO 7
J=J-K
K=K/2
GO TO 6
J=J+K
PI=3.14159265358979
DO 20 L=1,M
LE=2**L
LE1=LE/2
U=(1.0,0.0)
W=CMPLX(COS(PI/FLOAT(LE1)),SIN(PI/FLOAT(LE1)))
DO 10 J=1,LE1
DO 10 I=J,N,LE
IP=I+LE
T=X(IP)*U
X(IP)=X(I)-T
X(I)=X(IP)+T
U=U*W
RETURN
END

```

Fig. P6.4

Nella « subroutine » $\text{FFT}(X, M)$, X è un vettore complesso di dimensione N che contiene inizialmente la sequenza di ingresso $x(n)$ e alla fine contiene la trasformata $X(k)$. La quantità M è un intero, $M = \log_2 N$.

- Da una scorsa al programma indicate quali righe del testo sono dedicate a (1) inversione dei bit, (2) calcolo ricorsivo dei moltiplicatori esponenziali complessi, (3) calcolo delle farfalle.
 - Determinate su quale dei diagrammi di flusso del capitolo è basato il programma.
 - Sono stati inseriti tre errori nel programma qui riportato. Trovate questi errori e correggeteli opportunamente.
5. Il programma FORTRAN di fig. P6.5 è una realizzazione dell'algoritmo della decimazione in frequenza illustrato in fig. 6.18. Il programma calcola la DFT:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

Nella « subroutine » $\text{FFT}(X, M)$, X è un vettore complesso di dimensione N che contiene inizialmente la sequenza di ingresso $x(n)$ e alla fine contiene la trasformata $X(k)$. La quantità M è un intero, $M = \log_2 N$.

Questo programma è una realizzazione diretta del grafo di flusso di fig. 6.18. Il programma è molto elegante, ma non efficiente come potrebbe essere. Una maggior efficienza è ottenibile al prezzo di un programma più complesso.

Un significativo aumento di efficienza si ottiene notando che nell'ultimo stadio del grafo di flusso di fig. 6.18, i moltiplicatori complessi sono tutti unitari. Perciò, se l'ultimo stadio viene realizzato separatamente, possiamo eliminare $N/2$ moltiplicazioni complesse.

- Qual è la risultante riduzione percentuale del numero di moltiplicazioni?
- Modificate il programma per realizzare questo risparmio di moltiplicazioni.


```

SUBROUTINE FFT(X,M)          0001
COMPLEX X(1024),U,W,T       0002
N=2**M                       0003
PI=3.14159265358979         0004
DO 20 L=1,M                 0005
  LE=2**(M+1-L)              0006
  LE1=LE/2                   0007
  U=(1.0,0.0)                0008
  W=CMPLX(COS(PI/FLOAT(LE1)), -SIN(PI/FLOAT(LE1))) 0009
  DO 20 J=1,LE1              0010
    DO 10 I=J,N,LE           0011
      IP=I+LE1               0012
      T=X(I)+X(IP)          0013
      X(IP)=X(I)-X(IP)*U     0014
      X(I)=T                 0015
    U=U*W                   0016
  NV2=N/2                   0017
  NM1=N-1                   0018
  J=1                       0019
  DO 30 I=1,NM1             0020
    IF(I.GE.J) GO TO 25      0021
    T=X(J)                   0022
    X(J)=X(I)                0023
    X(I)=T                   0024
  K=NV2                     0025
  IF(K.GE.J) GO TO 30        0026
  J=J-K                     0027
  K=K/2                     0028
  GO TO 26                   0029
30 J=J+K                     0030
RETURN                       0031
END                           0032

```

Fig. P6.5

- (c) Molti piccoli calcolatori hanno compilatori FORTRAN che non prevedono aritmetica complessa. Modificate il programma dato in modo che si abbiano solo operazioni reali. Cioè, usando l'attuale « subroutine » come guida, scrivete una « subroutine »

FFT(XR, XI, M)

dove XR e XI sono vettori reali di dimensione N che inizialmente contengono la parte reale e la parte immaginaria dell'ingresso e alla fine la parte reale e immaginaria della trasformata.

6. Disegnare il diagramma di flusso per un algoritmo di FFT a decimazione nel tempo su nove (cioè, 3×3) punti.
7. Si supponga che N abbia un fattore 4; cioè, $N = 4 \cdot q_1$.
- (a) Esprimere la DFT $X(k)$ come combinazione di quattro DFT su q_1 punti (contrassegnare queste trasformate su q_1 punti con $G_l(k)$, $l = 0, 1, 2, 3$) come nelle (6.26) e (6.27).
- (b) Mostrare che il calcolo di base per valutare $X(k)$ dalle trasformate su q_1 punti $G_l(k)$ è come illustrato in fig. 6.28.
- (c) Mostrare che il grafo di flusso di fig. 6.28 può essere semplificato in quello di fig. 6.29.
- (d) Confrontare il numero di moltiplicazioni complesse necessarie per realizzare una FFT su 16 punti supponendo (1) che N sia scomposto come $N = 2 \cdot 2 \cdot 2 \cdot 2$, e (2) che N sia scomposto come $N = 4 \cdot 4$. (Supporre che i coefficienti che sono potenze intere di $W_{16}^4 = j$ non richiedano alcuna moltiplicazione in entrambi i calcoli).
8. Si supponga di disporre di un programma per calcolare una DFT

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

Mostrare come lo stesso programma possa essere usato per calcolare la DFT inversa

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j(2\pi/N)kn}, \quad n = 0, 1, \dots, N-1$$

9. In questo problema considereremo il procedimento per calcolare la DFT di quattro sequenze reali simmetriche o antisimmetriche di N punti dal calcolo di una DFT su N punti. Siano $x_1(n)$, $x_2(n)$, $x_3(n)$ e $x_4(n)$ le quattro sequenze reali di lunghezza N e $X_1(k)$, $X_2(k)$, $X_3(k)$, $X_4(k)$ le loro DFT. Supporremo dapprima che $x_1(n)$ e $x_2(n)$ siano simmetriche e che $x_3(n)$ ed $x_4(n)$ siano antisimmetriche; cioè,

$$x_1(n) = x_1((N-n))_N \mathcal{R}_N(n)$$

$$x_2(n) = x_2((N-n))_N \mathcal{R}_N(n)$$

$$x_3(n) = -x_3((N-n))_N \mathcal{R}_N(n)$$

$$x_4(n) = -x_4((N-n))_N \mathcal{R}_N(n)$$

- (a) Definiamo $y_1(n) = x_1(n) + x_3(n)$ e sia $Y_1(k)$ la DFT di $y_1(n)$. Determinare come $X_1(k)$ e $X_3(k)$ possano essere ricostruite da $Y_1(k)$.
- (b) $y_1(n)$ come definita nella parte (a) è reale. Analogamente, possiamo definire una sequenza reale $y_2(n) = x_2(n) + x_4(n)$. Sia $y_3(n)$ la sequenza complessa

$$y_3(n) = y_1(n) + jy_2(n)$$

Determinare dapprima come $Y_1(k)$ ed $Y_2(k)$ possano essere determinate da $Y_3(k)$ e quindi, usando i risultati della parte (a), mostrare come ottenere $X_1(k)$, $X_2(k)$, $X_3(k)$ e $X_4(k)$ da $Y_3(k)$.

Il risultato della parte (b) mostra che possiamo calcolare le DFT di quattro sequenze reali simultaneamente, se due sono simmetriche e due antisimmetriche.

Consideriamo ora il caso in cui tutte e quattro siano simmetriche; cioè,

$$x_i(n) = x_i((N-n))_N \mathcal{R}_N(n), \quad i = 1, 2, 3, 4$$

- (c) Si consideri una sequenza simmetrica reale $x_3(n)$. Mostrare che la sequenza $u_3(n) = x_3((n+1))_N - x_3((n-1))_N$ è una sequenza antisimmetrica; cioè

$$u_3(n) = -u_3((N-n))_N$$

- (d) Sia $U_3(k)$ la DFT di $u_3(n)$. Determinare $U_3(k)$ in termini di $X_3(k)$.
- (e) Usando il procedimento della parte (c), possiamo rappresentare la sequenza simmetrica $x_3(n) = x_1(n) + u_3(n)$. Determinare come $X_1(k)$ e $X_3(k)$ possono essere ricostruite da $Y_1(k)$.
- (f) Poniamo ora

$$y_3(n) = y_1(n) + jy_2(n)$$

dove

$$y_1(n) = x_1(n) + u_3(n)$$

$$y_2(n) = x_2(n) + u_4(n)$$

con

$$u_3(n) = [x_3((n+1))_N - x_3((n-1))_N] \mathcal{R}_N(n)$$

$$u_4(n) = [x_4((n+1))_N - x_4((n-1))_N] \mathcal{R}_N(n)$$

Determinare come ottenere $X_1(k)$, $X_2(k)$, $X_3(k)$ e $X_4(k)$ da $Y_3(k)$. [Notare che $X_3(k)$ e $X_4(k)$ non possono essere ottenute per $k = 0$, e $X_3(N/2)$ e $X_4(N/2)$ non possono essere determinate se N è pari].

- (g) $X_3(k)$ e $X_4(k)$ non possono essere determinate per $k = 0$ o $k = N/2$ usando il metodo precedente. Mostrare che questi punti possono essere valutati senza alcuna moltiplicazione.


```

SUBROUTINE FFT(X,M)
COMPLEX X(1024),U,W,T
N=2**M
PI=3.14159265358979
DO 20 L=1,M
LE=2** (M+1-L)
LE1=LE/2
U=(1.0,0.0)
W=CMPLX(COS(PI/FLOAT(LE1)), -SIN(PI/FLOAT(LE1)))
DO 20 J=1,LE1
DO 10 I=J,N,LE
IP=I+LE1
T=X(I)+X(IP)
X(IP)=(X(I)-X(IP))*U
X(I)=T
U=U*W
NV2=N/2
NM1=N-1
J=1
DO 30 I=1,NM1
IF(I.GE.J) GO TO 25
T=X(J)
X(J)=X(I)
X(I)=T
K=NV2
IF(K.GE.J) GO TO 30
J=J-K
K=K/2
GO TO 26
J=J+K
30 RETURN
END

```

Fig. P6.5

- (c) Molti piccoli calcolatori hanno compilatori FORTRAN che non prevedono aritmetica complessa. Modificate il programma dato in modo che si abbiano solo operazioni reali. Cioè, usando l'attuale « subroutine » come guida, scrivete una « subroutine »

FFT(XR, XI, M)

dove XR e XI sono vettori reali di dimensione N che inizialmente contengono la parte reale e la parte immaginaria dell'ingresso e alla fine la parte reale e immaginaria della trasformata.

6. Disegnare il diagramma di flusso per un algoritmo di FFT a decimazione nel tempo su nove (cioè, 3×3) punti.
7. Si supponga che N abbia un fattore 4; cioè, $N = 4 \cdot q$.
- Esprimere la DFT $X(k)$ come combinazione di quattro DFT su q punti (contrassegnare queste trasformate su q punti con $G_l(k)$, $l = 0, 1, 2, 3$) come nelle (6.26) e (6.27).
 - Mostrare che il calcolo di base per valutare $X(k)$ dalle trasformate su q punti $G_l(k)$ è come illustrato in fig. 6.28.
 - Mostrare che il grafo di flusso di fig. 6.28 può essere semplificato in quello di fig. 6.29.
 - Confrontare il numero di moltiplicazioni complesse necessarie per realizzare una FFT su 16 punti supponendo (1) che N sia scomposto come $N = 2 \cdot 2 \cdot 2 \cdot 2$, e (2) che N sia scomposto come $N = 4 \cdot 4$. (Supporre che i coefficienti che sono potenze intere di $W_{16}^4 = j$ non richiedano alcuna moltiplicazione in entrambi i calcoli).
8. Si supponga di disporre di un programma per calcolare una DFT

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

Mostrare come lo stesso programma possa essere usato per calcolare la DFT inversa

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j(2\pi/N)kn}, \quad n = 0, 1, \dots, N-1$$

9. In questo problema considereremo il procedimento per calcolare la DFT di quattro sequenze reali simmetriche o antisimmetriche di N punti dal calcolo di una DFT su N punti. Siano $x_1(n)$, $x_2(n)$, $x_3(n)$ e $x_4(n)$ le quattro sequenze reali di lunghezza N e $X_1(k)$, $X_2(k)$, $X_3(k)$, $X_4(k)$ le loro DFT. Supporremo dapprima che $x_1(n)$ e $x_2(n)$ siano simmetriche e che $x_3(n)$ ed $x_4(n)$ siano antisimmetriche; cioè,

$$x_1(n) = x_1((N-n))_N \mathcal{R}_N(n)$$

$$x_2(n) = x_2((N-n))_N \mathcal{R}_N(n)$$

$$x_3(n) = -x_3((N-n))_N \mathcal{R}_N(n)$$

$$x_4(n) = -x_4((N-n))_N \mathcal{R}_N(n)$$

- Definiamo $y_1(n) = x_1(n) + x_3(n)$ e sia $Y_1(k)$ la DFT di $y_1(n)$. Determinare come $X_1(k)$ e $X_3(k)$ possano essere ricostruite da $Y_1(k)$.
- $y_1(n)$ come definita nella parte (a) è reale. Analogamente, possiamo definire una sequenza reale $y_2(n) = x_2(n) + x_4(n)$. Sia $y_3(n)$ la sequenza complessa

$$y_3(n) = y_1(n) + jy_2(n)$$

Determinare dapprima come $Y_1(k)$ ed $Y_2(k)$ possano essere determinate da $Y_3(k)$ e quindi, usando i risultati della parte (a), mostrare come ottenere $X_1(k)$, $X_2(k)$, $X_3(k)$ e $X_4(k)$ da $Y_3(k)$.

Il risultato della parte (b) mostra che possiamo calcolare le DFT di quattro sequenze reali simultaneamente, se due sono simmetriche e due antisimmetriche.

Consideriamo ora il caso in cui tutte e quattro siano simmetriche; cioè,

$$x_i(n) = x_i((N-n))_N \mathcal{R}_N(n), \quad i = 1, 2, 3, 4$$

- Si consideri una sequenza simmetrica reale $x_3(n)$. Mostrare che la sequenza $u_3(n) = x_3((n+1))_N - x_3((n-1))_N$ è una sequenza antisimmetrica; cioè

$$u_3(n) = -u_3((N-n))_N$$

- Sia $U_3(k)$ la DFT di $u_3(n)$. Determinare $U_3(k)$ in termini di $X_3(k)$.
- Usando il procedimento della parte (c), possiamo rappresentare la sequenza simmetrica $x_3(n) = x_1(n) + u_3(n)$. Determinare come $X_1(k)$ e $X_3(k)$ possono essere ricostruite da $Y_1(k)$.
- Poniamo ora

$$y_3(n) = y_1(n) + jy_2(n)$$

dove

$$y_1(n) = x_1(n) + u_3(n)$$

$$y_2(n) = x_2(n) + u_4(n)$$

con

$$u_3(n) = [x_3((n+1))_N - x_3((n-1))_N] \mathcal{R}_N(n)$$

$$u_4(n) = [x_4((n+1))_N - x_4((n-1))_N] \mathcal{R}_N(n)$$

Determinare come ottenere $X_1(k)$, $X_2(k)$, $X_3(k)$ e $X_4(k)$ da $Y_3(k)$.

[Notare che $X_3(k)$ e $X_4(k)$ non possono essere ottenute per $k = 0$, e $X_3(N/2)$ e $X_4(N/2)$ non possono essere determinate se N è pari].

- $X_3(k)$ e $X_4(k)$ non possono essere determinate per $k = 0$ o $k = N/2$ usando il metodo precedente. Mostrare che questi punti possono essere valutati senza alcuna moltiplicazione.

10. Nel calcolare la DFT di sequenze reali, è possibile ridurre l'ammontare dei calcoli utilizzando il fatto che la sequenza è reale. In questo problema discuteremo diversi modi di ottenere tale riduzione.

(a) Sia $x(n)$ una sequenza a valori reali con N punti e sia $X(k)$ la sua DFT, con parte reale e immaginaria indicate da $X_R(k)$ e $X_I(k)$, rispettivamente, così che

$$X(k) = X_R(k) + jX_I(k)$$

Mostrare che se $x(n)$ è reale, allora $X_R(k)$ è pari a $X_I(k)$ è dispari, cioè

$$X_R(k) = X_R((N-k))_N \quad \text{e} \quad X_I(k) = -X_I((N-k))_N$$

(b) Si considerino due sequenze a valori reali $x_1(n)$ e $x_2(n)$ con DFT $X_1(k)$ e $X_2(k)$, rispettivamente. Sia $g(n)$ una sequenza a valori complessi, definita come $g(n) = x_1(n) + jx_2(n)$, e sia $G(k)$ la sua DFT. Siano $G_{OR}(k)$, $G_{ER}(k)$, $G_{OI}(k)$ e $G_{EI}(k)$, rispettivamente, la parte dispari della parte reale, la parte pari della parte reale, la parte dispari della parte immaginaria e la parte pari della parte immaginaria. Determinare $X_1(k)$ ed $X_2(k)$ in termini di $G_{OR}(k)$, $G_{ER}(k)$, $G_{OI}(k)$ e $G_{EI}(k)$.

Il risultato ricavato nella parte (b) può essere utilizzato in molti modi: se si hanno due sequenze reali di cui si vuole la DFT, possiamo calcolare le loro trasformate contemporaneamente e quindi separare le trasformate usando il risultato della parte (b). Un'altra possibilità, che considereremo ora, è di spezzare inizialmente una sequenza reale in due sequenze più brevi, e quindi mettere insieme questi risultati per ottenere la DFT della sequenza totale.

(c) Supponiamo che $x(n)$ sia una sequenza a valori reali con N punti e che N sia divisibile per 2. Siano $x_1(n)$ e $x_2(n)$ le due sequenze di $N/2$ punti definite da

$$x_1(n) = x(2n), \quad n = 0, 1, 2, \dots, N/2 - 1$$

$$x_2(n) = x(2n+1), \quad n = 0, 1, 2, \dots, N/2 - 1$$

Determinare $X(k)$ in termini di $X_1(k)$ e $X_2(k)$.

11. Si consideri una sequenza di lunghezza finita $x(n)$ tale che $x(n) = 0$ per $n < n_0$ ed $n > N-1+n_0$. Supponiamo di voler calcolare campioni della trasformata z di $x(n)$ nei seguenti punti del piano z :

$$z_k = re^{j(\theta + (2\pi/M)k)}, \quad k = 0, 1, \dots, M-1$$

dove $M < N$.

Descrivere dettagliatamente un procedimento efficiente per calcolare $X(z)$ nei punti desiderati.

12. Si consideri una sequenza $x(n)$ di durata finita lunga M tale che $x(n) = 0$ per $n < 0$ ed $n \geq M$. Vogliamo calcolare campioni della trasformata z

$$X(z) = \sum_{n=0}^{M-1} x(n)z^{-n}$$

in N punti equispaziati sulla circonferenza unitaria, cioè in

$$z = e^{j(2\pi/N)k}, \quad k = 0, 1, \dots, N-1$$

Determinare e giustificare procedimenti per calcolare gli N campioni di $X(z)$ usando solo una DFT su N punti per i casi

(a) $N \leq M$.

(b) $N > M$.

13. $X(e^{j\omega})$ indica la trasformata di Fourier di una sequenza di durata finita $x(n)$ di lunghezza 10. Desideriamo calcolare 10 campioni di $X(e^{j\omega})$ a frequenze $\omega_k = (2\pi k^2/100)$, $k = 0, 1, \dots, 9$, senza calcolare più campioni di $X(e^{j\omega})$ di quelli richiesti e poi non considerarne alcuni. Discutere la possibilità di ottenere tale risultato con ciascuno dei seguenti metodi:

(a) Direttamente, usando un algoritmo di FFT su 10 punti.

(b) Usando l'algoritmo della trasformata z chirp.

(c) Usando l'algoritmo di Goertzel.

14. Un'applicazione dell'algoritmo della CZT è quella di mettere in evidenza i picchi di risonanza in uno spettro. In generale, se calcoliamo la trasformata di una sequenza su un percorso nel piano z vicino ad un polo, ci aspettiamo di osservare una risonanza. Nell'applicare l'algoritmo della CZT, o nel calcolare la DFT, la sequenza che si sta analizzando deve essere di durata finita. Se non lo è, la sequenza deve prima essere troncata. Mentre la trasformata della sequenza originaria aveva dei poli, la trasformata della sequenza troncata può avere solo zeri (tranne che in $z = 0$ o $z = \infty$). Lo scopo di questo problema è mostrare che nella trasformata della sequenza di lunghezza finita si osserverà ancora una risposta di tipo risonante.

(a) Sia $x(n) = u(n)$. Disegnare il diagramma di poli e zeri della sua trasformata z , $X(z)$.

(b) Sia

$$\hat{x}(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{altrove} \end{cases}$$

cioè $\hat{x}(n)$ è uguale ad $x(n)$ troncata dopo N punti. Disegnare la configurazione di zeri e poli di $\hat{X}(z)$, la trasformata z di $\hat{x}(n)$.

(c) Disegnare $|\hat{X}(e^{j\omega})|$ in funzione di ω . Indicare nel disegno l'effetto che si ha aumentando N .

15. Scegliere il finale corretto della frase seguente. La trasformata z chirp può essere usata per calcolare la trasformata z , $H(z)$, di una sequenza di durata finita $h(n)$ in punti $\{z_k\}$ sull'asse reale del piano z tali che

(a) $z_k = a^k$, $k = 0, 1, \dots, N-1$ per a reale, $a \neq \pm 1$.

(b) $z_k = ak$, $k = 0, 1, \dots, N-1$ per a reale, $a \neq 0$.

(c) Entrambe le parti (a) e (b).

(d) Né la parte (a) né la parte (b), cioè la CZT non può essere usata per calcolare valori di $H(z)$ per z reale.

16. Bluestein [14] ha proposto uno schema ricorsivo per calcolare tutti i valori della DFT

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn}, \quad k = 0, 1, \dots, N-1$$

per N quadrato perfetto, cioè $N = M^2$.

(a) Usando la sostituzione

$$-nk = \frac{(k-n)^2}{2} - \frac{n^2}{2} - \frac{k^2}{2}$$

mostrare che $X(k)$ può essere espressa come la convoluzione

$$X(k) = h^*(k) \sum_{n=0}^{N-1} (x(n)h^*(n))h(k-n)$$

dove

$$h(n) = e^{j(\pi/N)n^2}, \quad -\infty < n < \infty$$

(b) Mostrare che i valori desiderati di $X(k)$ (cioè, per $k = 0, 1, \dots, N-1$) possono anche essere ottenuti valutando la convoluzione della parte (a) per $k = N, N+1, \dots, 2N-1$.

(c) Usare il risultato della parte (b) per mostrare che $X(k)$ è anche uguale all'uscita del sistema di fig. P6.16 per $k = N, N+1, \dots, 2N-1$, dove $\hat{h}(k)$ è la sequenza di durata finita

$$\hat{h}(k) = e^{j(\pi/N)k^2}, \quad 0 \leq k \leq 2N-1$$

- (d) Usando il fatto che $N = M^2$, mostrare che la funzione caratteristica corrispondente alla risposta all'impulso $\hat{h}(k)$ è

$$\begin{aligned}\hat{H}(z) &= \sum_{k=0}^{2N-1} e^{j(\pi k^2/N)} z^{-k} \\ &= \sum_{r=0}^{M-1} z^{-r} e^{j(\pi r^2/N)} \frac{1 - z^{-2M}}{1 + e^{j(2\pi/M)r} z^{-M}}\end{aligned}$$

(Suggerimento: Esprimere k come $k = r + IM$).

- (e) La precedente espressione per $\hat{H}(z)$ suggerisce una realizzazione ricorsiva del sistema a risposta all'impulso di durata finita. Disegnare un diagramma di flusso di tale realizzazione.
- (f) Usare il risultato della parte (e) per determinare il numero totale di moltiplicazioni e somme complesse richieste per calcolare tutti gli N valori desiderati di $X(k)$. Confrontare col numero richiesto per la valutazione diretta di $X(k)$.

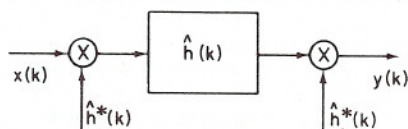


Fig. P6.16

7. TRASFORMATE DI HILBERT DISCRETE

7.0 INTRODUZIONE

In quasi tutti i settori dove vengono usate le tecniche di Fourier per rappresentare ed analizzare processi fisici, si presentano casi in cui esistono relazioni precise tra la parte reale e quella immaginaria o tra il modulo e la fase della trasformata di Fourier. Queste relazioni vanno sotto nomi diversi a seconda del settore d'interesse, ma spesso (e il campo dell'elaborazione numerica dei segnali non fa eccezione) vengono indicate come *trasformate di Hilbert*.

Vedremo, ad esempio, che se una sequenza è causale, allora la parte reale e la parte immaginaria della sua trasformata di Fourier sono legate da una trasformata di Hilbert integrale. In questo capitolo deriveremo un certo numero di relazioni di questo tipo, che sono importanti sia nella teoria che nelle applicazioni dell'elaborazione numerica dei segnali. Per esempio, si vedrà nel cap. 10 che una comprensione approfondita di molti degli argomenti di questo capitolo è condizione essenziale per capire appieno i concetti legati alla deconvoluzione omomorfica.

In generale, le funzioni complesse che si impiegano per la rappresentazione matematica di segnali e sistemi a tempo discreto sono, come suol dirsi, « ben comportate ». Le trasformate z di cui ci siamo interessati avevano, con poche eccezioni, regioni ben definite dove la serie di potenze era assolutamente convergente. Poiché una serie di potenze rappresenta una funzione analitica dentro la sua regione di convergenza [1,2], ne consegue che le trasformate z sono funzioni analitiche all'interno delle loro regioni di convergenza. Per definizione di funzione analitica, ciò significa che la trasformata z ha derivata unica in ogni punto interno alla regione di convergenza. Inoltre, l'analiticità implica che la trasformata z e tutte le sue derivate sono funzioni continue all'interno della regione di convergenza.

Queste proprietà delle funzioni analitiche comportano dei vincoli piuttosto pesanti sul comportamento della trasformata z nella sua regione di convergenza. Uno di tali vincoli è che la parte reale e la parte immaginaria soddisfano le condizioni di Cauchy-Riemann, che mettono in relazione le derivate parziali delle parti reale e immaginaria di una funzione analitica. Un altro vincolo è il teorema integrale di Cauchy, per mezzo del quale si può esprimere il valore di una funzione complessa in qualunque punto interno alla sua regione di analiticità, in termini dei valori della

funzione sul contorno. Sulla base di queste relazioni valide per funzioni analitiche, è possibile, sotto certe condizioni, derivare delle relazioni integrali esplicite tra la parte reale e quella immaginaria di una trasformata z su un percorso chiuso interno alla regione di convergenza. Nella letteratura matematica queste relazioni vengono spesso chiamate *formule di Poisson* [2,3]. Nell'ambito della teoria dei sistemi esse sono note come *trasformate di Hilbert* ed hanno tradizionalmente avuto un ruolo importante nell'elaborazione dei segnali, sia dal punto di vista teorico che pratico [4,5].

Anche se le trasformate di Hilbert possono essere sviluppate in modo formale a partire dalle proprietà delle funzioni analitiche (v. probl. 2 e 4), in questo capitolo useremo un approccio in qualche misura più intuitivo. In particolare, esse saranno sviluppate partendo dalla considerazione che le parti reale e immaginaria (sul circolo unitario) della trasformata z di una sequenza causale sono le trasformate delle componenti pari e dispari della sequenza. Come si vedrà, una sequenza causale ha la proprietà di essere completamente determinata dalla sua parte pari, il che implica che la trasformata z della sequenza originale è specificata completamente dalla sua parte reale sul circolo unitario. Oltre ad applicare questa proprietà per specificare la trasformata z di una sequenza causale in termini della sua parte reale sul circolo unitario, la si può anche usare, sotto certe condizioni, per esprimere la trasformata z di una sequenza in termini del suo *modulo* sul circolo unitario.

Un concetto importante nell'elaborazione dei segnali a tempo continuo è poi quello di segnale analitico [6]. Un segnale analitico è una funzione complessa (che è analitica) del tempo, avente trasformata di Fourier nulla per frequenze negative. Una *sequenza complessa* non può, formalmente, essere considerata analitica, poiché è una funzione di variabile intera. Tuttavia è possibile, in maniera analoga a quanto fatto sopra, porre in relazione le parti reale e immaginaria di una sequenza complessa il cui spettro sia nullo sul circolo unitario per $-\pi < \omega < 0$. Una strada simile si può anche seguire per legare le parti reale e immaginaria della trasformata di Fourier discreta per una sequenza periodica oppure, cosa equivalente, di durata finita. In questo caso la condizione di « causalità » diventa che la sequenza periodica sia nulla nella seconda metà di ogni periodo.

Pertanto, nella discussione che segue, si applicherà il concetto di causalità per mettere in relazione le componenti pari e dispari di una funzione o, in maniera equivalente, le parti reale e immaginaria delle sue trasformate. I casi trattati sono quattro. Nel primo si legano le parti reale e immaginaria della trasformata di Fourier $H(e^{j\omega})$ di una sequenza $h(n)$ che è nulla per $n < 0$. Nel secondo si mettono in relazione le parti reale e immaginaria del *logaritmo* della trasformata di Fourier, con la condizione che la trasformata inversa del *logaritmo* della trasformata sia zero per $n < 0$. Mettere in relazione le parti reale e immaginaria del *logaritmo* della trasformata corrisponde a mettere in relazione il *logaritmo* del modulo e la fase di $H(e^{j\omega})$. Nel terzo caso si svilupperà una relazione tra le parti

reale e immaginaria della DFT per sequenze periodiche ovvero, equivalentemente, per sequenze di durata finita considerate lunghe N ma con gli ultimi $N/2$ campioni nulli. Infine, metteremo in relazione le parti reale e immaginaria di una sequenza complessa la cui trasformata di Fourier, considerata come funzione periodica di ω , sia nulla nella seconda metà di ogni periodo.

7.1 SUFFICIENZA DELLA SOLA PARTE REALE O IMMAGINARIA PER SEQUENZE CAUSALI

Qualsiasi sequenza può essere espressa come somma di una sequenza pari e di una dispari. Più precisamente, se $h_e(n)$ e $h_o(n)$ indicano le parti pari e dispari di $h(n)$, allora risulta

$$h(n) = h_e(n) + h_o(n) \quad (7.1)$$

dove è

$$h_e(n) = \frac{1}{2}[h(n) + h(-n)] \quad (7.2)$$

e

$$h_o(n) = \frac{1}{2}[h(n) - h(-n)] \quad (7.3)$$

Le espressioni scritte sopra valgono per una sequenza arbitraria, sia essa causale o no, reale o meno. Se però $h(n)$ è causale, allora si può riottenere da $h_e(n)$ o, per $n \neq 0$, da $h_o(n)$. Si consideri, ad esempio, la sequenza causale $h(n)$ e le sue componenti pari e dispari mostrate in fig. 7.1. Poiché $h(n)$ è causale, cioè $h(n)$ è nulla per $n < 0$ e quindi $h(-n)$ è nulla per $n > 0$, non vi è sovrapposizione tra le parti diverse da zero di $h(n)$ e $h(-n)$, eccetto in $n = 0$.

Dovrebbe essere chiaro dalla fig. 7.1 e dalle relazioni (7.2) e (7.3) che per sequenze causali risulta

$$h(n) = \begin{cases} 2h_e(n), & n > 0 \\ h_e(n), & n = 0 \\ 0, & n < 0 \end{cases} \quad (7.4)$$

e

$$h(n) = \begin{cases} 2h_o(n), & n > 0 \\ 0, & n < 0 \end{cases} \quad (7.5)$$

In maniera equivalente, se definiamo la sequenza

$$u_+(n) = \begin{cases} 2, & n > 0 \\ 1, & n = 0 \\ 0, & n < 0 \end{cases} \quad (7.6)$$

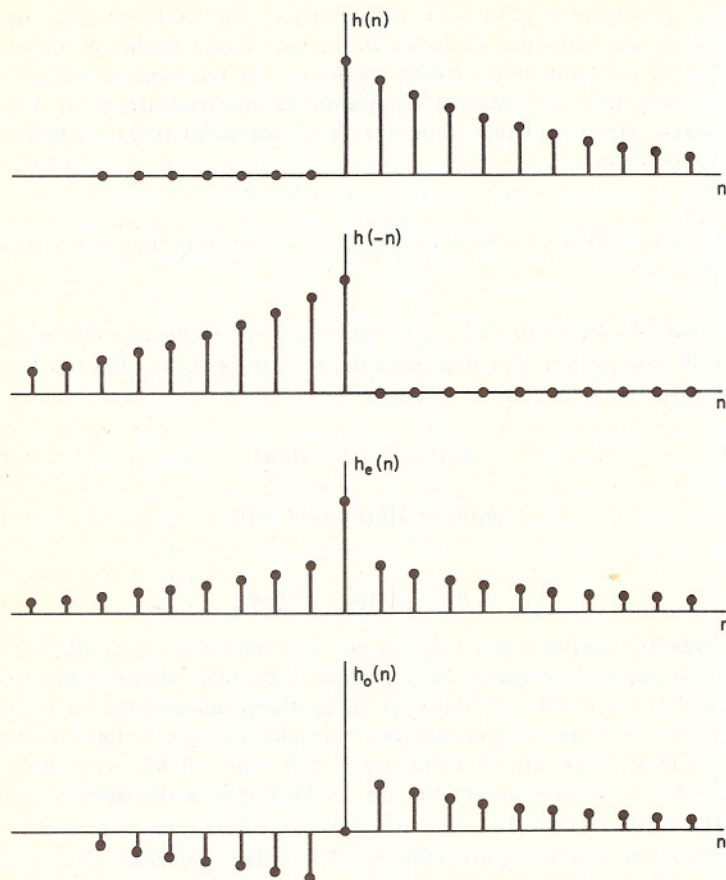


Fig. 7.1 Parte pari e parte dispari di una sequenza reale causale.

allora si può scrivere

$$h(n) = h_e(n)u_+(n) \quad (7.7)$$

e

$$h(n) = h_o(n)u_+(n) + h(0)\delta(n) \quad (7.8)$$

Notiamo che $h(n)$ può essere ricostruita completamente da $h_e(n)$. D'altra parte, $h_o(n)$ è sempre zero per $n = 0$, e di conseguenza $h(n)$ può essere riottenuta da $h_o(n)$ solo per $n \neq 0$.

Una conseguenza importante delle relazioni (7.7) e (7.8) è che la trasformata di Fourier di una sequenza reale, causale e stabile

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega})$$

è completamente nota se ne conosciamo o la parte reale $H_R(e^{j\omega})$ oppure la parte immaginaria $H_I(e^{j\omega})$ insieme ad $h(0)$. Questo è vero perché $H_R(e^{j\omega})$

è la trasformata di Fourier di $h_e(n)$ e $jH_I(e^{j\omega})$ è la trasformata di Fourier di $h_o(n)$. Per esempio, possiamo calcolare $h_e(n)$ da $H_R(e^{j\omega})$ e poi, usando la (7.7), possiamo ottenere $h(n)$, da cui si può risalire a $H(e^{j\omega})$.

Più in generale, si può dimostrare che se $h(n)$ è reale, causale e stabile, allora $H(z)$ può essere determinata ovunque all'esterno del cerchio unitario (cioè nella sua regione di convergenza) dalla conoscenza di $H_R(e^{j\omega})$ oppure di $H_I(e^{j\omega})$ insieme ad $h(0)$ [7-9]. Considerando dunque $H(z)$ al di fuori del cerchio unitario, cioè per $z = re^{j\omega}$ con $r > 1$, si ha

$$H(z)|_{z=re^{j\omega}} = H(re^{j\omega}) = \sum_{n=0}^{\infty} h(n)r^{-n}e^{-j\omega n}$$

ovvero, usando la relazione (7.7),

$$H(re^{j\omega}) = \sum_{n=-\infty}^{\infty} h_e(n)u_+(n)r^{-n}e^{-j\omega n}$$

Questa espressione può anche essere interpretata come la trasformata di Fourier del prodotto $h_e(n) \cdot [r^{-n}u_+(n)]$. Quindi $H(re^{j\omega})$ può essere ricavata come convoluzione della trasformata di Fourier di $h_e(n)$ e della trasformata di Fourier della sequenza $r^{-n}u_+(n)$. La trasformata di Fourier di $h_e(n)$ è $H_R(e^{j\omega})$ e, se $r > 1$, quella di $r^{-n}u_+(n)$ è $(1 + r^{-1}e^{-j\omega}) / (1 - r^{-1}e^{-j\omega})$ (si noti che la trasformata di Fourier di $r^{-n}u_+(n)$ non esiste in senso stretto se $r = 1$). Ora, usando il teorema della convoluzione complessa del par. 2.3.9, si ottiene la relazione

$$H(z)|_{z=re^{j\omega}} = \frac{1}{2\pi j} \oint_C \frac{H_R(v)(e^{j\omega} + r^{-1}v) dv}{(e^{j\omega} - r^{-1}v)v} \quad (7.9a)$$

o, in alternativa,

$$H(z) = \frac{1}{2\pi j} \oint_C \frac{H_R(v)(z + v) dv}{(z - v)v}, \quad |z| > 1 \quad (7.9b)$$

Negli integrali su percorso chiuso delle (7.9a) e (7.9b), C deve essere la circonferenza unitaria, in quanto si assume che è noto solamente $H_R(e^{j\omega})$. L'utilità di queste relazioni si dimostra in particolare quando $H_R(e^{j\omega})$ può essere espressa come funzione razionale di $e^{j\omega}$, poiché in questo caso l'integrale può essere valutato facilmente usando il calcolo dei residui.

ESEMPIO. Supponiamo sia dato

$$H_R(e^{j\omega}) = \frac{1 - \alpha \cos \omega}{1 - 2\alpha \cos \omega + \alpha^2}, \quad |\alpha| < 1$$

Troviamo $H(z)$ usando le (7.9). Innanzitutto scriviamo $H_R(e^{j\omega})$ come funzione razionale di $e^{j\omega}$,

$$H_R(e^{j\omega}) = \frac{1 - \alpha(e^{j\omega} + e^{-j\omega})/2}{(1 - \alpha e^{-j\omega})(1 - \alpha e^{j\omega})}$$

Sostituiamo poi questa espressione nella (7.9b), ottenendo

$$H(z) = \frac{1}{2\pi j} \oint_C \frac{(1 - \alpha(v + v^{-1})/2) z + v dv}{(1 - \alpha v^{-1})(1 - \alpha v) z - v v}$$

dove C è il circolo unitario. Riscrivendo questa espressione in modo da evidenziare i poli dell'integrando, si ha

$$H(z) = \frac{1}{2\pi j} \oint_C \frac{(v - \alpha(v^2 + 1)/2)(z + v) dv}{(v - \alpha)(1 - \alpha v)(z - v)v}$$

I poli dell'integrando sono mostrati in fig. 7.2, da cui si nota che soltanto i poli in

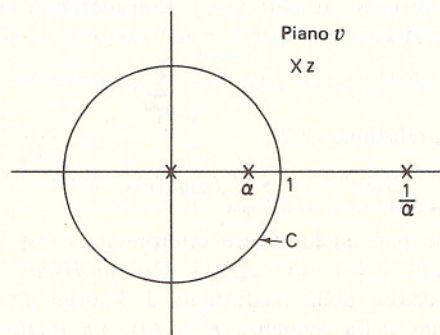


Fig. 7.2 Posizioni dei poli nell'esempio di calcolo della trasformata z usando l'integrazione su percorso chiuso

$v = 0$ e $v = \alpha$ sono dentro il cerchio unitario. Perciò, usando il teorema dei residui si ha

$$\begin{aligned} H(z) &= \frac{-(\alpha/2)z}{-\alpha z} + \frac{(\alpha - \alpha(\alpha^2 + 1)/2)(z + \alpha)}{(1 - \alpha^2)(z - \alpha)\alpha} \\ &= \frac{1}{2} + \frac{1}{2} \frac{z + \alpha}{z - \alpha} = \frac{z}{z - \alpha} \end{aligned}$$

Questa espressione è stata ricavata assumendo che sia $|z| > 1$; notiamo però che la regione di analiticità è $|z| > \alpha$. Abbiamo quindi ottenuto la trasformata z direttamente dalla sua parte reale sul circolo unitario.

Le relazioni (7.9a) e (7.9b) esprimono $H(z)$ al di fuori del circolo unitario in termini della sua parte reale sul circolo medesimo. Tuttavia, è anche utile scrivere questa espressione come integrale su percorso non chiuso. Poniamo quindi $v = e^{j\theta}$ nella (7.9a), in modo da ottenere

$$\begin{aligned} H(z) \Big|_{z=re^{j\omega}} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} H_R(e^{j\theta}) P_r(\theta - \omega) d\theta \\ &+ \frac{j}{2\pi} \int_{-\pi}^{\pi} H_R(e^{j\theta}) Q_r(\theta - \omega) d\theta \end{aligned} \quad (7.10)$$

dove abbiamo definito

$$P_r(\theta) = \operatorname{Re} \left(\frac{1 + r^{-1}e^{j\theta}}{1 - r^{-1}e^{j\theta}} \right) = \frac{1 - r^{-2}}{1 - 2r^{-1} \cos \theta + r^{-2}} \quad (7.11a)$$

e

$$Q_r(\theta) = \operatorname{Im} \left(\frac{1 + r^{-1}e^{j\theta}}{1 - r^{-1}e^{j\theta}} \right) = \frac{2r^{-1} \sin \theta}{1 - 2r^{-1} \cos \theta + r^{-2}} \quad (7.11b)$$

Le funzioni $P_r(\theta)$ e $Q_r(\theta)$ vengono spesso chiamate, rispettivamente, il *nucleo di Poisson* e il *nucleo di Poisson coniugato* [2,3]. Uguagliando, nella (7.10), le parti reale e immaginaria, si ottiene

$$H_R(re^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_R(e^{j\theta}) P_r(\theta - \omega) d\theta \quad (7.12)$$

e

$$H_I(re^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_R(e^{j\theta}) Q_r(\theta - \omega) d\theta \quad (7.13)$$

Come si vede, abbiamo ricavato relazioni integrali reali per le parti reale e immaginaria della trasformata z all'esterno del circolo unitario in termini della sola parte reale valutata su di esso.

Con passaggi matematici simili si arriva, partendo dalla relazione (7.8), alla rappresentazione con integrale su percorso chiuso

$$H(z) \Big|_{z=re^{j\omega}} = \frac{1}{2\pi} \oint_C \frac{H_I(v)(e^{j\omega} + r^{-1}v) dv}{(e^{j\omega} - r^{-1}v)v} + h(0) \quad (7.14a)$$

oppure

$$H(z) = \frac{1}{2\pi} \oint_C \frac{H_I(v)(z + v) dv}{(z - v)v} + h(0), \quad |z| > 1 \quad (7.14b)$$

dove il percorso C è ancora il circolo unitario. Trasformando la (7.14a) in un'espressione contenente integrali su percorso non chiuso e uguagliando le parti reale e immaginaria, si giunge a

$$H_R(re^{j\omega}) = -\frac{1}{2\pi} \int_{-\pi}^{\pi} H_I(e^{j\theta}) Q_r(\theta - \omega) d\theta + h(0) \quad (7.15)$$

e

$$H_I(re^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_I(e^{j\theta}) P_r(\theta - \omega) d\theta \quad (7.16)$$

dove $P_r(\theta)$ e $Q_r(\theta)$ sono quelle delle espressioni (7.11a) e (7.11b).

Per ottenere delle relazioni dirette tra la parte reale e la parte immaginaria valutate sul circolo unitario, occorre fare il limite delle espressioni (7.13) e (7.15) per r che tende a uno. Questo è vero purché si esegua l'operazione di integrazione per prima. Se invece cerchiamo di ottenere una relazione diretta per $H_I(e^{j\omega})$ in termini di $H_R(e^{j\omega})$, scambiando l'ordine delle operazioni di integrazione e di limite, ci troviamo di fronte ad un integrale improprio, in quanto è

$$\lim_{r \rightarrow 1} Q_r(\theta) = \frac{2 \sin \theta}{2(1 - \cos \theta)} = \cot \left(\frac{\theta}{2} \right)$$

e la funzione $\cot(\theta/2)$ ha una singolarità in $\theta = 0$. Possiamo ricavare le relazioni cercate se valutiamo con attenzione gli integrali impropri in prossimità dei punti singolari dell'integrando. Questo può essere fatto formal-

mente interpretando gli integrali come *valori principali di Cauchy* [10]. In questo modo la (7.13) diventa

$$H_I(e^{j\omega}) = \frac{1}{2\pi} P \int_{-\pi}^{\pi} H_R(e^{j\theta}) \cot\left(\frac{\theta - \omega}{2}\right) d\theta \quad (7.17)$$

e la (7.15) diventa

$$H_R(e^{j\omega}) = h(0) - \frac{1}{2\pi} P \int_{-\pi}^{\pi} H_I(e^{j\theta}) \cot\left(\frac{\theta - \omega}{2}\right) d\theta \quad (7.18)$$

dove la lettera P sta a significare valore principale di Cauchy. Il significato di valore principale di Cauchy, con riferimento, per es., alla (7.17), è chiarito nella (7.19):

$$H_I(e^{j\omega}) = \frac{1}{2\pi} \lim_{\epsilon \rightarrow 0} \left\{ \int_{\omega+\epsilon}^{\pi} H_R(e^{j\theta}) \cot\left(\frac{\theta - \omega}{2}\right) d\theta + \int_{-\pi}^{\omega-\epsilon} H_R(e^{j\theta}) \cot\left(\frac{\theta - \omega}{2}\right) d\theta \right\} \quad (7.19)$$

Notiamo che $H_I(e^{j\omega})$ è ottenuto per mezzo della convoluzione periodica di $\cot(-\omega/2)$ con $H_R(e^{j\omega})$, e che occorre fare particolare attenzione in prossimità della singolarità per $\theta = \omega$. In modo analogo, l'espressione (7.18) contiene la convoluzione periodica di $\cot(-\omega/2)$ con $H_I(e^{j\omega})$.

Le due funzioni che compaiono nell'integrale di convoluzione della (7.17) ovvero della (7.19) sono mostrate nella fig. 7.3. L'esistenza del limite nella (7.19) è assicurata dal fatto che la funzione $\cot[(\theta - \omega)/2]$ è antisimmetrica nel punto singolare ($\theta = \omega$) e l'intervallo è collocato in modo simmetrico rispetto alla singolarità.

La valutazione degli integrali nelle espressioni precedenti è ulteriormente complicata quando $H(z)$ ha poli sul circolo unitario. Nella nostra discussione abbiamo assunto che il circolo unitario sia contenuto intera-

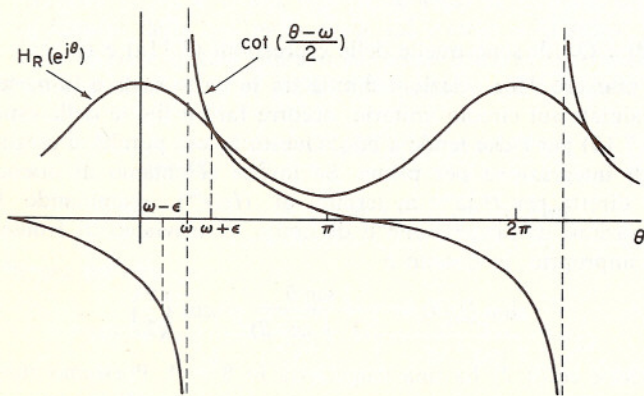


Fig. 7.3 Interpretazione della trasformata di Hilbert come convoluzione periodica

mente nella regione di convergenza di $H(z)$, e che perciò $H(z)$ non abbia poli su di esso. Si può risolvere il problema dei poli sul circolo unitario ammettendo impulsi nella trasformata di Fourier oppure usando un percorso opportuno (che aggiri i poli) negli integrali. Siccome però la giustificazione matematica di queste procedure ci porterebbe troppo lontano, trascureremo ogni approfondimento di questo aspetto.

7.2 CONDIZIONE DI FASE MINIMA

Nel paragrafo precedente è stata ricavata la trasformata z di una sequenza causale a partire dalla sua parte reale o immaginaria sul circolo unitario. Discuteremo in questo paragrafo le condizioni sotto le quali la trasformata z si può ottenere a partire dal suo modulo o dalla sua fase sul circolo unitario. Queste condizioni sono di notevole importanza in molti casi sia teorici che pratici. Ad esempio, le specifiche sui filtri numerici sono spesso date in termini di modulo della risposta in frequenza. In questi casi la risposta di fase non può essere scelta arbitrariamente se è richiesto un sistema stabile e causale. I risultati di questo paragrafo sono anche molto importanti nella teoria dei sistemi omorfi, che verrà sviluppata nel cap. 10. Un altro esempio si ha nella teoria e applicazione del filtraggio inverso, dove occorre ottenere una curva di fase opportuna, essendo data solo una funzione di autocorrelazione (o, cosa equivalente, il modulo quadrato della trasformata di Fourier) [9, 11-13].

Supponiamo che $H(z)$ sia rappresentata in forma polare (modulo e fase) come

$$H(z) = |H(z)| e^{j\arg[H(z)]}$$

Consideriamo poi il logaritmo complesso di $H(z)$, definito come

$$\hat{H}(z) = \log[H(z)] = \log|H(z)| + j\arg[H(z)] \quad (7.20)$$

Se pensiamo $\hat{H}(z)$ come la trasformata z di una sequenza $\hat{h}(n)$, allora i risultati del paragrafo precedente implicano che $\log|H(e^{j\omega})|$ e $\arg[H(e^{j\omega})]$ sono la trasformata di Hilbert uno dell'altra se e solo se $\hat{h}(n)$ è una sequenza reale, causale e stabile. Nel trattare questo argomento occorre fare molta attenzione nell'interpretare la definizione (7.20). In particolare, il logaritmo di zero diverge e la definizione di $\arg[H(z)]$ è ambigua in quanto si può aggiungere qualsiasi multiplo di 2π alla fase senza alterare il valore di $H(z)$. Poiché intendiamo interpretare $\hat{H}(z)$ come la trasformata z di una sequenza reale, causale e stabile, desideriamo che la regione di convergenza di $\hat{H}(z)$ comprenda il circolo unitario, e di conseguenza è richiesto che $\hat{H}(z)$ sia analitica in una regione che lo include. All'interno di questa regione, allora, $\hat{H}(z)$ deve avere una rappresentazione in serie di potenze convergente del tipo

$$\hat{H}(z) = \sum_{n=0}^{\infty} \hat{h}(n) z^{-n}, \quad R_{h-} < |z|$$

dove è $R_{h-} < 1$. Poiché $\hat{H}(z)$ va all'infinito in corrispondenza sia dei poli che degli zeri di $H(z)$, dobbiamo imporre la condizione che non vi siano né poli né zeri di $H(z)$ dentro la regione di convergenza associata ad $\hat{H}(z)$. Nonostante $\arg[H(z)]$ non sia in generale unico, l'ambiguità è risolta dal fatto che l'analiticità di $\hat{H}(z)$ implica che le sue parti reale e immaginaria devono essere funzioni continue di z : di conseguenza, se $\hat{H}(z)$ deve essere analitica, occorre definire $\arg[H(z)]$ nella (7.20) in modo che sia una funzione continua. Inoltre, porremo il vincolo che quando $h(n)$ è reale $\hat{H}(z)$ sia la trasformata z di una sequenza reale. Di conseguenza, $\arg[H(z)]$ sarà definito in modo tale che per $z = e^{j\omega}$ sia una funzione dispari e continua di ω ¹.

Consideriamo ora una sequenza $\hat{h}(n)$ reale e stabile la cui trasformata z sia $\hat{H}(z)$. Dal paragrafo precedente dovrebbe essere chiaro che se $\hat{h}(n)$ è causale, allora $\hat{H}(z)$ e quindi $H(z)$ possono essere ricostruite a partire da $\hat{H}_R(e^{j\omega}) = \log|H(e^{j\omega})|$ o da $\hat{H}_I(e^{j\omega}) = \arg[H(e^{j\omega})]$. In modo equivalente, se $\hat{h}(n)$ è reale, stabile e causale, si possono applicare le (7.17) e (7.18) per porre in relazione il logaritmo del modulo e la fase di $H(e^{j\omega})$ come segue

$$\log|H(e^{j\omega})| = \hat{h}(0) - \frac{1}{2\pi} P \int_{-\pi}^{\pi} \arg[H(e^{j\theta})] \cot\left(\frac{\theta - \omega}{2}\right) d\theta \quad (7.21)$$

$$\arg[H(e^{j\omega})] = \frac{1}{2\pi} P \int_{-\pi}^{\pi} \log|H(e^{j\theta})| \cot\left(\frac{\theta - \omega}{2}\right) d\theta \quad (7.22)$$

Si osservi che se $\hat{h}(0)$ non è noto, $|H(e^{j\omega})|$ è determinato da $\arg[H(e^{j\omega})]$ solo a meno di una costante moltiplicativa.

Il vincolo che $\log|H(e^{j\omega})|$ ed $\arg[H(e^{j\omega})]$ siano legate dalla trasformata di Hilbert è spesso indicato come *condizione di fase minima* [4, 5, 14]². Ciò corrisponde alla proprietà che la sequenza $\hat{h}(n)$ sia causale. Allora, come è stato discusso nel cap. 2, $H(z)$ deve essere analitica in una regione $|z| > R_{h-}$ con $R_{h-} < 1$; in altri termini, $\hat{H}(z)$ deve essere analitica ovunque fuori del cerchio unitario. Perciò non possono esistere singolarità di $\hat{H}(z)$ all'esterno del cerchio unitario. Poiché poi è $\hat{H}(z) = \log H(z)$, questo implica che non possono esistere né poli né zeri di $H(z)$ fuori del cerchio unitario. Tale vincolo su $H(z)$ può essere visto come un'espressione alternativa della condizione di fase minima. Una condizione equivalente è che esista un *sistema inverso* con funzione di trasferimento $H^{-1}(z)$ tale che

$$H^{-1}(z)H(z) = 1$$

Poiché $H^{-1}(z) = 1/H(z)$, è chiaro che $H(z)$ deve avere tutti i suoi poli e zeri dentro il cerchio unitario affinché esista un sistema inverso stabile e causale.

Da questo momento useremo il termine *sistema a fase minima* per indicare un sistema la cui risposta in frequenza è a fase minima, cioè con logaritmo del modulo e fase che sono trasformate di Hilbert uno dell'altra.

¹ Si noti che assumiamo che sia $H(e^{j\omega}) > 0$ per $\omega = 0$.

² Il motivo del termine « fase minima » sarà chiarito nella discussione che segue.

Analogamente, una *sequenza a fase minima* è una sequenza la cui trasformata di Fourier è a fase minima. Occorre sottolineare a questo punto che un sistema (o una sequenza) può essere causale ma non a fase minima. Tuttavia, tutti i sistemi (sequenze) stabili e a fase minima sono causali.

Per capire la relazione tra la causalità di $\hat{h}(n)$ e la posizione dei poli e degli zeri di $H(z)$, è istruttivo cercare un procedimento per ottenere $\hat{h}(n)$. In particolare, sappiamo dal cap. 2 che $-z[d\hat{H}(z)/dz]$ è la trasformata z di $n\hat{h}(n)$. Ma risulta

$$-z \frac{d\hat{H}(z)}{dz} = -z \frac{d}{dz} [\log H(z)] = \frac{-z}{H(z)} \frac{dH(z)}{dz} \quad (7.23)$$

Se $H(z)$ è una funzione razionale di z , $\hat{H}(z)$ non è razionale ma la sua derivata sì, e di conseguenza può essere caratterizzata in termini di poli e zeri. Esprimendo $H(z)$ come un rapporto di polinomi

$$H(z) = \frac{P(z)}{Q(z)}$$

otteniamo

$$\frac{-z}{H(z)} \frac{dH(z)}{dz} = \frac{-z \left[Q(z) \frac{dP(z)}{dz} - P(z) \frac{dQ(z)}{dz} \right]}{P(z)Q(z)}$$

Notiamo perciò che i poli della derivata di $\hat{H}(z)$ sono le radici di $P(z)Q(z)$, cioè i poli e gli zeri di $H(z)$. Poiché consideriamo il cerchio unitario come interno alla regione di convergenza, $n\hat{h}(n)$ ovvero $\hat{h}(n)$ sarà causale se e solo se tutti i poli e gli zeri di $H(z)$ sono interni al cerchio unitario³.

ESEMPIO. Si consideri la sequenza $h(n) = \alpha^n u(n)$, per la quale è $H(z) = 1/(1 - \alpha z^{-1})$, con $|\alpha| < 1$. $H(z)$ ha uno zero in $z = 0$ e un polo in $z = \alpha$. Poiché $|\alpha| < 1$, tutti i poli e gli zeri sono interni al cerchio unitario, e di conseguenza $h(n)$ è a fase minima. Per verificare che $h(n)$ è davvero causale, calcoliamola usando la relazione (7.23). Risulta in questo caso

$$\frac{-z}{H(z)} \frac{dH(z)}{dz} = \frac{\alpha}{z - \alpha} = \frac{\alpha z^{-1}}{1 - \alpha z^{-1}}$$

Perciò, siccome assumiamo che il cerchio unitario sia dentro la regione di convergenza, si ha

$$n\hat{h}(n) = \alpha^n u(n - 1)$$

e quindi $\hat{h}(n)$ è causale.

La sequenza $\hat{h}(n)$ giocherà un ruolo particolarmente importante nel cap. 10. Non considereremo qui ulteriormente le proprietà della sequenza $\hat{h}(n)$ per concentrarci invece sulle proprietà delle sequenze a fase minima.

Una sequenza a fase minima ha la proprietà che tutti i poli e gli zeri della sua trasformata z si trovano dentro il cerchio unitario. In gene-

³ Compresi i poli e gli zeri all'infinito; cioè, se $H(z)$ è a fase minima, $\lim_{z \rightarrow \infty} H(z)$ deve essere una costante di valore finito e diverso da 0.

rale, un sistema stabile e causale ha tutti i poli all'interno del circolo unitario, ma questo non vale necessariamente per gli zeri. Dimosteremo adesso che qualunque sistema può essere rappresentato come la cascata di un sistema a fase minima e di un sistema passa-tutto, essendo quest'ultimo definito come un sistema per cui il modulo della funzione di trasferimento vale uno a tutte le frequenze. Perciò, se $H_{ap}(z)$ indica la trasformata z di un sistema passa-tutto, $|H_{ap}(e^{j\omega})| = 1$ per ogni ω .

La funzione di trasferimento di un semplice sistema passa-tutto del primo ordine è

$$H_{ap}(z) = \frac{z^{-1} - a}{1 - az^{-1}} \quad (7.24)$$

Il fatto che il modulo della $H_{ap}(e^{j\omega})$ espressa nella (7.24) sia unitario è esaminato nel probl. 6 di questo capitolo. Il diagramma di poli e zeri corrispondente è mostrato in fig. 7.4 con $0 < a < 1$. Più in generale, le funzioni di trasferimento razionali di sistemi passa-tutto sono espresse come prodotto di fattori della forma

$$\frac{z^{-1} - a^*}{1 - az^{-1}}$$

e di conseguenza hanno la proprietà che i loro poli e zeri sono coniugati reciproci.

Consideriamo un sistema $H(z)$ a fase non minima, con, ad esempio, uno zero esterno al cerchio unitario in $z = 1/z_0$, $|z_0| < 1$, e gli altri poli e zeri dentro il cerchio unitario. Allora $H(z)$ può essere espressa come

$$H(z) = H_1(z)(z^{-1} - z_0) \quad (7.25)$$

dove $H_1(z)$ è a fase minima. Possiamo esprimere la (7.25) in maniera equivalente come

$$\begin{aligned} H(z) &= H_1(z)(z^{-1} - z_0) \frac{1 - z_0^* z^{-1}}{1 - z_0^* z^{-1}} \\ &= H_1(z)(1 - z_0^* z^{-1}) \frac{z^{-1} - z_0}{1 - z_0^* z^{-1}} = H_{min}(z) \frac{z^{-1} - z_0}{1 - z_0^* z^{-1}} \end{aligned}$$

Poiché $|z_0| < 1$, il fattore $H_1(z) (1 - z_0^* z^{-1})$ è a fase minima e il fattore $(z^{-1} - z_0)/(1 - z_0^* z^{-1})$ è passa-tutto. Il termine $H_{min}(z) = H_1(z) (1 - z_0^* z^{-1})$ differisce da $H(z)$ in quanto lo zero di $H(z)$ che era fuori del cerchio unitario in $z = 1/z_0$ risulta per $H_{min}(z)$ ribaltato all'interno del cerchio unitario in $z = z_0^*$. È chiaro che questo esempio può essere generalizzato a tutti i sistemi a fase non minima con funzione di trasferimento razionale. Concludiamo perciò che qualsiasi funzione di trasferimento $H(z)$ razionale corrispondente a un sistema causale può essere espressa nella forma

$$H(z) = H_{min}(z)H_{ap}(z) \quad (7.26)$$

dove $H_{min}(z)$ è a fase minima e $H_{ap}(z)$ è passa-tutto. Tutti i poli e gli zeri di $H(z)$ interni al cerchio unitario appaiono anche in $H_{min}(z)$. I poli e gli

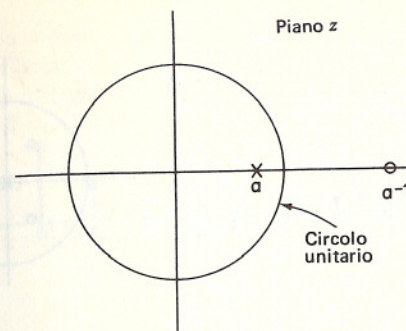


Fig. 7.4 Posizioni di poli e zeri per un sistema passa-tutto del primo ordine

zeri di $H(z)$ esterni al cerchio unitario compaiono in $H_{min}(z)$ in posizione coniugata reciproca, cioè ribaltati rispetto alla circonferenza unitaria. Possiamo quindi costruire un sistema a fase minima a partire da un sistema a fase non minima, mantenendo inalterato il modulo della funzione di trasferimento, ribaltando all'interno del cerchio unitario quegli zeri che erano all'esterno. Dualmente, data una funzione di trasferimento a fase minima, possiamo ottenere un sistema a fase non minima ribaltando gli zeri fuori del cerchio unitario. Ad esempio, nel caso di sequenze di lunghezza finita, la trasformata z è semplicemente un polinomio in z^{-1} e $H(z)$ ha poli solo in $z = 0$. Per una sequenza lunga M , $H(z)$ ha $M-1$ zeri. Per una risposta di modulo dato, possiamo avere fino a 2^{M-1} diverse curve di fase semplicemente ribaltando gli zeri rispetto alla circonferenza unitaria.

ESEMPIO. Consideriamo una risposta all'impulso di durata finita e a fase minima, lunga $N = 5$ campioni. La risposta all'impulso di questo sistema è illustrata in fig. 7.5(a). La funzione di trasferimento corrispondente è

$$H_{min}(z) = \frac{1}{r^2} (1 - re^{j\theta} z^{-1})^2 (1 - re^{-j\theta} z^{-1})^2 \quad (7.27)$$

con $r = 0.55$ e $\theta = 2\pi/3$. Le funzioni logaritmo del modulo e fase della risposta in frequenza sono rappresentate in fig. 7.5(c) e (d) rispettivamente (si noti che $\arg[H_{min}(e^{j\omega})]$ è rappresentato modulo 2π e $\log|H_{min}(e^{j\omega})|$ è normalizzato a un valore di picco di 0 dB per comodità di rappresentazione). In base alla discussione precedente, possiamo ottenere un nuovo sistema avente lo stesso modulo della risposta moltiplicando $H_{min}(z)$ per un'opportuna funzione di trasferimento passa-tutto come risulta dalla relazione (7.26). In questo caso, possiamo ribaltare all'esterno una coppia degli zeri complessi coniugati usando il sistema passa-tutto

$$H_{ap}(z) = \frac{z^{-1} - re^{-j\theta}}{1 - re^{j\theta} z^{-1}} \frac{z^{-1} - re^{j\theta}}{1 - re^{-j\theta} z^{-1}} \quad (7.28)$$

Otteniamo quindi

$$\begin{aligned} H(z) &= H_{min}(z)H_{ap}(z) \\ &= (1 - re^{j\theta} z^{-1})(1 - re^{-j\theta} z^{-1})(1 - r^{-1}e^{j\theta} z^{-1})(1 - r^{-1}e^{-j\theta} z^{-1}) \end{aligned} \quad (7.29)$$

Notiamo che i quattro zeri di $H(z)$ hanno la simmetria coniugata reciproca che è una proprietà caratteristica dei sistemi a fase lineare. In effetti, si vede in fig. 7.6(a) che la risposta all'impulso $h(n)$ associata a $H(z)$ è simmetrica attorno a $n = 2$, cosa che implica una fase lineare con pendenza corrispondente a un ritardo di due campioni.

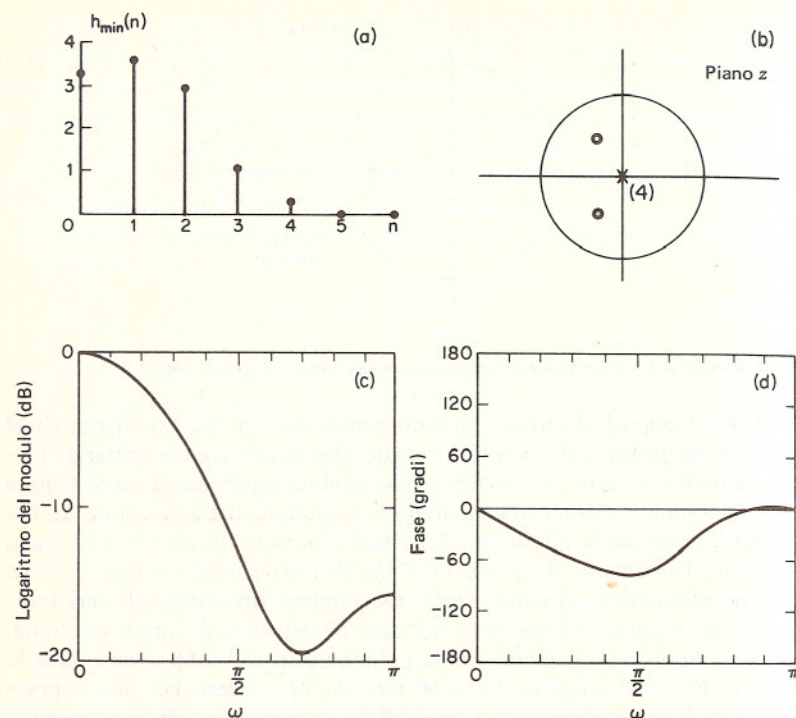


Fig. 7.5 Esempio di sistema a fase minima: (a) risposta all'impulso; (b) diagramma di poli e zeri nel piano z ; (c) $20 \log_{10} |H_{\min}(e^{j\omega})|$; (d) $\arg[H_{\min}(e^{j\omega})]$

Come si vede dal confronto di fig. 7.5(c) e di fig. 7.6(c), $|H(e^{j\omega})|$ è identico a $|H_{\min}(e^{j\omega})|$; tuttavia, la risposta all'impulso e le risposte di fase corrispondenti a $H_{\min}(z)$ e $H(z)$ sono decisamente diverse.

La fig. 7.7 mostra (a) il diagramma di poli e zeri nel piano z e (b) $\arg[H_{ap}(e^{j\omega})]$ per il sistema passa-tutto. Il modulo di $H_{ap}(e^{j\omega})$ è, naturalmente, unitario per tutti i valori di ω . Abbiamo rappresentato ancora $\arg[H_{ap}(e^{j\omega})]$ modulo 2π semplicemente per comodità. È però chiaro dalla fig. 7.7(b) che se la fase viene calcolata come funzione continua di ω , allora $\arg[H_{ap}(e^{j\omega})]$ è sempre negativo. Quando questa curva di fase viene sommata a quella del sistema a fase minima [fig. 7.5(d)], ne risulta la fase lineare di fig. 7.6(d).

Questo semplice esempio serve a illustrare alcune importanti proprietà generali dei sistemi a fase minima, che vale la pena di sottolineare. Innanzitutto, un confronto tra le curve di fase delle fig. 7.5(d) e 7.6(d) chiarisce la ragione dell'espressione « a fase minima ». Come già discusso, se consideriamo l'insieme delle sequenze reali, causali e stabili, aventi tutte la stessa risposta d'ampiezza, allora le trasformate z di tutte queste sequenze possono essere espresse, in base alla (7.26), come il prodotto di una trasformata z a fase minima e di una funzione passa-tutto. Come si è visto nell'esempio precedente e come sarà discusso nel probl. 8 di questo capitolo, la funzione passa-tutto ha fase negativa per $0 < \omega < \pi$, e, di conseguenza,

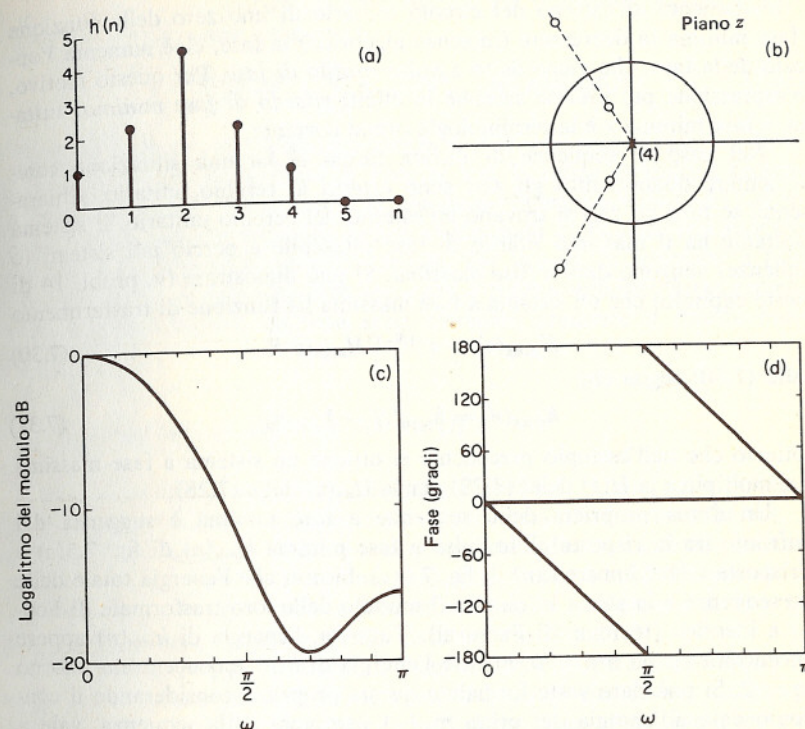


Fig. 7.6 Sistema a fase lineare: (a) risposta all'impulso; (b) diagramma di poli e zeri nel piano z ; (c) $20 \log_{10} |H(e^{j\omega})|$; (d) $\arg[H(e^{j\omega})]$ [il modulo della risposta è identico a quello di fig. 7.5(c)]

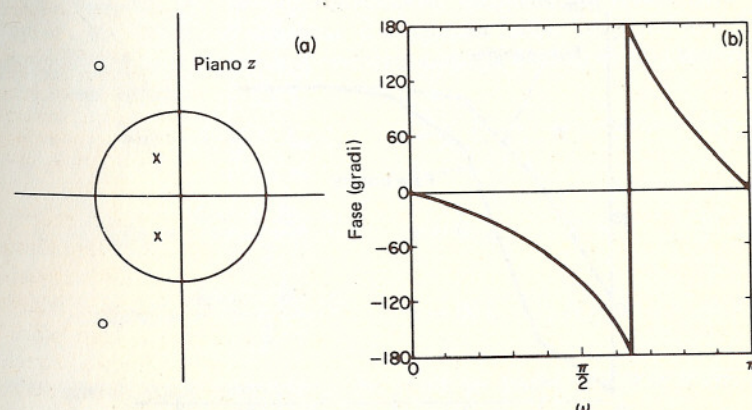


Fig. 7.7 Sistema passa-tutto per ottenere la fig. 7.6 dalla 7.5: (a) diagramma di poli e zeri nel piano z ; (b) $\arg[H_{ap}(e^{j\omega})]$

il ribaltamento all'esterno del cerchio unitario di uno zero della funzione a fase minima fa decrescere (in senso algebrico) la fase, cioè aumenta l'opposto della fase, che viene detto a volte *ritardo di fase*. Per questo motivo, un'espressione più precisa sarebbe in effetti *ritardo di fase minimo*; tuttavia, « fase minima » è la terminologia ormai corrente.

Nel caso di sequenze di durata finita, si ha una situazione complementare quando tutti gli zeri sono *esterni* al cerchio unitario. Chiaramente, se tutti gli zeri si trovano all'esterno del cerchio unitario, il sistema risultante ha il massimo ritardo di fase ottenibile e perciò tali sistemi (o sequenze) vengono detti *a fase massima*. Si può dimostrare (v. probl. 14 di questo capitolo) che un sistema a fase massima ha funzione di trasferimento

$$H_{\max}(z) = z^{-(N-1)} H_{\min}(z^{-1}) \quad (7.30)$$

Dalla (7.30) segue che

$$h_{\max}(n) = h_{\min}(N-1-n) \quad (7.31)$$

Notiamo che nell'esempio precedente si ottiene un sistema a fase massima se si moltiplica la $H(z)$ della (7.29) per la $H_{ap}(z)$ della (7.28).

Un'ultima proprietà delle sequenze a fase minima è suggerita dal confronto tra la risposta all'impulso a fase minima $h_{\min}(n)$ di fig. 7.5(a) e la risposta a fase lineare $h(n)$ di fig. 7.6(a). Si noti che l'energia totale delle due sequenze è la stessa in quanto il modulo delle loro trasformate di Fourier è identico (teorema di Parseval). Tuttavia, l'energia di $h_{\min}(n)$ appare concentrata vicino a $n=0$, mentre l'energia di $h(n)$ è concentrata attorno a $n=2$. Si può dare veste formale a questa proprietà considerando il contributo dato all'energia dai primi $m+1$ campioni della sequenza, vale a dire

$$E(m) = \sum_{n=0}^m |h(n)|^2 \quad (7.32)$$

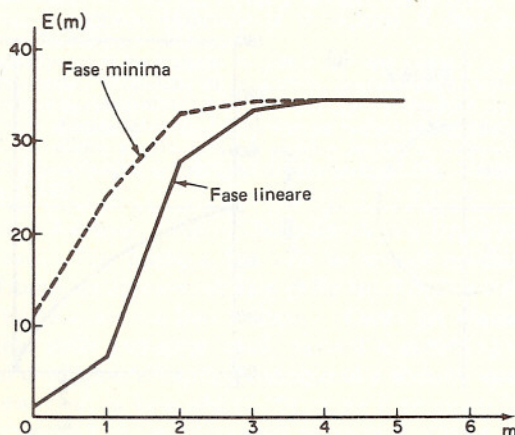


Fig. 7.8 Concentrazione dell'energia per due risposte all'impulso con trasformate di Fourier aventi lo stesso modulo

Questa quantità è rappresentata nella fig. 7.8 per le risposte $h_{\min}(n)$ e $h(n)$ dell'esempio precedente. Si noti che

$$\sum_{n=0}^m |h(n)|^2 \leq \sum_{n=0}^m |h_{\min}(n)|^2, \quad \text{per qualsiasi } m \quad (7.33)$$

Una dimostrazione che la disuguaglianza (7.33) vale in generale per tutte le sequenze aventi trasformata di Fourier con lo stesso modulo è presentata nel probl. 11 di questo capitolo. Possiamo interpretare la (7.33) affermando che di tutte le sequenze aventi trasformata di Fourier con modulo identico, $h_{\min}(n)$ è quella meno ritardata. Per questo, le sequenze a fase minima vengono spesso chiamate *sequenze a ritardo minimo*. Analogamente, le sequenze a fase massima sono chiamate *sequenze a ritardo massimo* [12,13].

7.3 TRASFORMATE DI HILBERT PER LA DFT

Si è visto che le sequenze periodiche e quelle di durata finita possono essere rappresentate in termini della trasformata di Fourier discreta. I risultati del paragrafo precedente non si applicano direttamente alla trasformata di Fourier discreta. Possiamo però, con un'opportuna definizione di causalità, legare le parti reale e immaginaria della trasformata di Fourier discreta in un modo simile a quello visto nel par. 7.1 [7,8].

Per arrivare a queste relazioni, è opportuno considerare una sequenza periodica $\tilde{h}(n)$ di periodo N . Si ricorderà dal cap. 3 che questo nostro approccio, anche se riferito a sequenze periodiche, vale altrettanto bene per sequenze di lunghezza finita pur di interpretare tutti gli indici modulo N . E invero, anche se la nostra trattazione riguarderà le proprietà delle rappresentazioni in serie di Fourier discreta (DFS), vedremo che i risultati valgono direttamente per le rappresentazioni mediante la trasformata di Fourier discreta (DFT) di sequenze di lunghezza finita. Come nel par. 7.2, la sequenza $\tilde{h}(n)$ può essere rappresentata come la somma di una sequenza pari e di una dispari

$$\tilde{h}(n) = \tilde{h}_e(n) + \tilde{h}_o(n), \quad n = 0, 1, \dots, N-1 \quad (7.34)$$

dove è

$$\tilde{h}_e(n) = \frac{\tilde{h}(n) + \tilde{h}(-n)}{2}, \quad n = 0, 1, \dots, N-1 \quad (7.35a)$$

e

$$\tilde{h}_o(n) = \frac{\tilde{h}(n) - \tilde{h}(-n)}{2}, \quad n = 0, 1, \dots, N-1 \quad (7.35b)$$

In tutto quanto segue assumeremo che N sia un intero pari. Per N dispari si ottengono risultati simili, anche se non identici.

Una sequenza periodica non può, ovviamente, essere causale nel senso usato nel par. 7.1. Definiremo tuttavia sequenza periodica « causale »

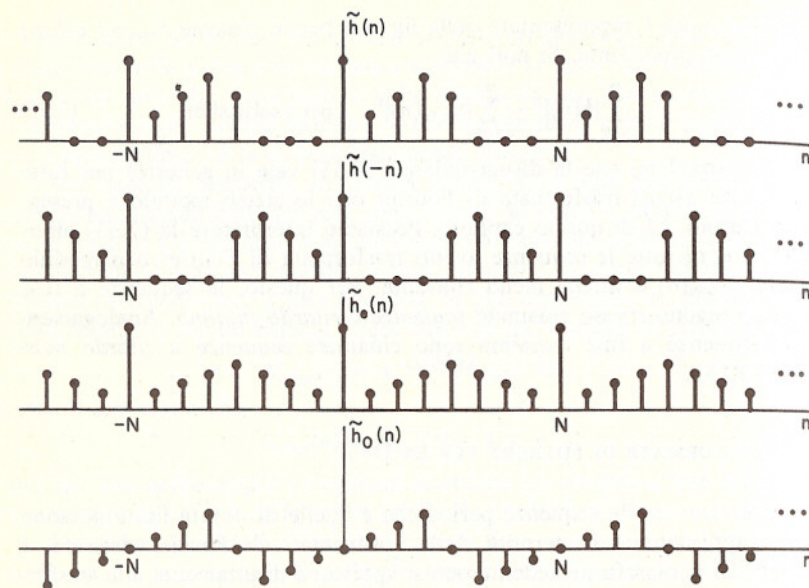


Fig. 7.9 Parte pari e dispari di una sequenza reale, periodica e « causale »

quella per cui $\tilde{h}(n)=0$ per $N/2 < n < N$. In altri termini, $\tilde{h}(n)$ è identicamente nulla nella seconda metà del periodo. Assumiamo che N sia pari: il caso di N dispari è considerato nel probl. 16 di questo capitolo. Si noti anche che per la periodicità di $\tilde{h}(n)$ risulta $\tilde{h}(n)=0$ per $-N/2 < n < 0$. Per sequenze di lunghezza finita questa limitazione significa che la sequenza è considerata lunga N , quando in effetti la seconda metà dei punti è nulla. La fig. 7.9 mostra un esempio di una sequenza periodica e delle sue parti pari e dispari per $N=8$. Poiché $\tilde{h}(n)$ è zero nella seconda metà di ogni periodo, $\tilde{h}(-n)$ vale zero nella prima metà di ogni periodo e, di conseguenza, non vi è sovrapposizione (eccetto che per $n=0$ e $n=N/2$) tra le parti diverse da zero di $\tilde{h}(n)$ e $\tilde{h}(-n)$. Per questa ragione dovrebbe essere chiaro che per sequenze periodiche « causali » risulta

$$\tilde{h}(n) = \begin{cases} 2\tilde{h}_e(n), & n = 1, 2, \dots, (N/2) - 1 \\ \tilde{h}_e(n), & n = 0, N/2 \\ 0, & n = (N/2 + 1), \dots, N - 1 \end{cases}$$

e

$$\tilde{h}(n) = \begin{cases} 2\tilde{h}_o(n), & n = 1, 2, \dots, (N/2) - 1 \\ 0, & n = (N/2 + 1), \dots, N - 1 \end{cases}$$

In modo equivalente, se definiamo la sequenza periodica $\tilde{u}_N(n)$

$$\tilde{u}_N(n) = \begin{cases} 1, & n = 0, N/2 \\ 2, & n = 1, 2, \dots, (N/2) - 1 \\ 0, & n = (N/2 + 1), \dots, N - 1 \end{cases}$$

allora, per N pari, $\tilde{h}(n)$ può essere espresso come

$$\tilde{h}(n) = \tilde{h}_e(n)\tilde{u}_N(n) \quad (7.36)$$

e

$$\tilde{h}(n) = \tilde{h}_o(n)\tilde{u}_N(n) + h(0)\delta(n) + h\left(\frac{N}{2}\right)\delta\left(n - \frac{N}{2}\right) \quad (7.37)$$

Notiamo che $\tilde{h}(n)$ può essere riottenuto completamente da $\tilde{h}_e(n)$. D'altro canto, $\tilde{h}_o(n)$ è sempre nullo per $n=0$ e $n=N/2$, per cui $\tilde{h}(n)$ può essere ricavato da $\tilde{h}_o(n)$ solo per $n \neq 0$ o $n \neq N/2$.

Abbiamo visto nel cap. 3 che per una sequenza reale periodica di periodo N con serie di Fourier discreta $\tilde{H}(k)$, la parte reale di $\tilde{H}(k)$, $\tilde{H}_R(k)$, è la DFS di $\tilde{h}_e(n)$ e $j\tilde{H}_I(k)$ è la DFS di $\tilde{h}_o(n)$. Perciò una conseguenza importante delle (7.36) e (7.37) è che queste implicano che per una sequenza periodica (oppure di lunghezza finita) di periodo N , che sia causale nel senso prima definito, $\tilde{H}(k)$ può essere ricostruita dalla sua parte reale o (quasi interamente) dalla sua parte immaginaria. In modo equivalente, $\tilde{H}_I(k)$ può essere ricostruita da $\tilde{H}_R(k)$ e $\tilde{H}_R(k)$ da $\tilde{H}_I(k)$.

Specificamente, la DFS della sequenza $\tilde{u}_N(n)$ è

$$\tilde{U}_N(k) = \begin{cases} N, & k = 0 \\ -j2 \cot\left(\frac{\pi}{N}k\right), & k \text{ dispari} \\ 0, & k \text{ pari} \end{cases} \quad (7.38)$$

Dalla (7.36) notiamo che la DFS di $\tilde{h}(n)$ è la convoluzione circolare di $\tilde{H}_R(k)$ con $\tilde{U}_N(k)$. Perciò risulta

$$\begin{aligned} \tilde{H}(k) &= \tilde{H}_R(k) + j\tilde{H}_I(k) \\ &= \frac{1}{N} \sum_{m=0}^{N-1} \tilde{H}_R(m) \tilde{U}_N(k-m) \\ &= \tilde{H}_R(k) + \frac{1}{N} \sum_{m=0}^{N-1} \tilde{H}_R(m) \tilde{V}_N(k-m) \end{aligned}$$

dove è

$$\tilde{V}_N(k) = \tilde{U}_N(k) - N\delta(k) = \begin{cases} -j2 \cot(\pi k/N), & k \text{ dispari} \\ 0, & k \text{ pari} \end{cases}$$

Uguagliando le parti reale e immaginaria si ottiene

$$j\tilde{H}_I(k) = \frac{1}{N} \sum_{m=0}^{N-1} \tilde{H}_R(m) \tilde{V}_N(k-m) \quad (7.39a)$$

Analogamente, partendo dalla relazione (7.37) si può dimostrare che risulta

$$\tilde{H}_R(k) = \frac{1}{N} \sum_{m=0}^{N-1} j\tilde{H}_I(m) \tilde{V}_N(k-m) + \tilde{h}(0) + \tilde{h}(N/2)(-1)^k \quad (7.39b)$$

I secondi membri delle uguaglianze (7.39a) e (7.39b) sono convoluzioni circolari e possono essere valutate usando la DFS. Ad esempio, la (7.39a) può

essere valutata calcolando dapprima la DFS inversa di $\tilde{H}_R(k)$, che fornisce $\tilde{h}_e(n)$. Moltiplicando $\tilde{h}_e(n)$ per $\tilde{u}_N(n)$ e calcolando la DFS si ricava poi la trasformata completa $\tilde{H}(k)$.

Nel par. 3.5 si è vista l'utilità di introdurre una notazione particolare per facilitare l'interpretazione delle espressioni della DFS nel contesto delle sequenze di lunghezza finita. Una sequenza di durata finita $h(n)$ è quindi pensata come un periodo di una sequenza periodica $\tilde{h}(n)$, vale a dire

$$h(n) = \tilde{h}(n)\mathcal{R}_N(n)$$

dove è

$$\begin{aligned}\mathcal{R}_N(n) &= 1 & 0 \leq n \leq N-1 \\ &= 0 & \text{altrove}\end{aligned}$$

In alternativa, si ottiene la sequenza periodica $\tilde{h}(n)$ interpretando l'indice n modulo N . A questo scopo, abbiamo introdotto la notazione

$$\tilde{h}(n) = h((n))_N$$

Queste convenzioni sono state anche applicate all'espressione della DFS $\tilde{H}(k)$ per ottenere l'espressione della DFT $H(k)$.

Usando questa notazione, la relazione (7.39a) si può riscrivere come

$$\begin{aligned}jH_I(k) &= \frac{1}{N} \sum_{m=0}^{N-1} H_R(m)V_N((k-m))_N & 0 \leq k \leq N-1 \\ &= 0 & \text{altrove}\end{aligned} \quad (7.40a)$$

e la (7.39b) come

$$\begin{aligned}H_R(k) &= \frac{1}{N} \sum_{m=0}^{N-1} jH_I(m)V_N((k-m))_N \\ &\quad + h(0) + (-1)^k h(N/2) & 0 \leq k \leq N-1 \\ &= 0 & \text{altrove}\end{aligned} \quad (7.40b)$$

dove

$$\begin{aligned}V_N(k) &= -j2 \cot(\pi k/N) & 0 < k < N-1, k \text{ dispari} \\ &= 0 & \text{altrove}\end{aligned}$$

Definizioni simili possono naturalmente essere date per u_N e la sua DFT $U_N(k)$.

Quando abbiamo discusso la sufficienza della parte reale per la conoscenza di tutta la trasformata z , abbiamo anche applicato le proprietà allora stabilite per porre in relazione il logaritmo del modulo e la fase di sequenze a fase minima. Per la trasformata di Fourier discreta non è possibile in generale sviluppare una relazione analoga, in base alla quale legare il logaritmo del modulo e la fase della DFT. Il motivo di questa impossibilità è che la discussione che abbiamo fatto vale per sequenze di lunghezza finita, per cui la trasformata z ha solo zeri. Tuttavia, il logaritmo di una trasformata $H(z)$ ha singolarità in corrispondenza sia dei poli che degli zeri di $H(z)$, e perciò la sua trasformata z inversa è di durata infinita. Di

conseguenza, la trasformata inversa del logaritmo della trasformata non può essere rappresentata, in generale, con una trasformata di Fourier discreta.

Naturalmente, è possibile costruire una funzione di fase a partire dal logaritmo del modulo di una DFT usando il procedimento descritto prima, vale a dire, calcolare la DFT inversa di $\log |H(k)|$, moltiplicare per $u_N(n)$ e calcolare la DFT della sequenza risultante. La parte reale del risultato è $\log |H(k)|$, e la parte immaginaria è un'approssimazione della fase minima.

Per capire questo procedimento, assumiamo che $H(z)$ sia la trasformata z di una sequenza $h(n)$ di lunghezza finita. Se $H(z)$ non ha zeri esterni al cerchio unitario, si può calcolare $\arg[H(e^{j\omega})]$ conoscendo solo $\log |H(e^{j\omega})|$. Inoltre

$$\hat{H}(z) = \log [H(z)]$$

corrisponde ad una sequenza causale $\tilde{h}(n)$, che in generale sarà di durata infinita. La DFT di $h(n)$ è

$$H(k) = H(z)|_{z=e^{j(2\pi k/N)}}, \quad k = 0, 1, \dots, N-1$$

dove N è scelto grande almeno quanto la lunghezza della sequenza $h(n)$. La DFT

$$\hat{H}_p(k) = \log [H(k)] = \log |H(k)| + j \arg [H(k)]$$

corrisponde ad una sequenza con *aliasing*

$$\hat{h}_p(n) = \sum_{r=-\infty}^{\infty} \hat{h}(n + rN)$$

È chiaro che quanto più grande scegliamo N , tanto migliore è il risultato del procedimento di moltiplicare $\hat{h}_p(n)$ per $u_N(n)$. Il calcolo della DFT per ottenere $\log |H(k)|$ per la parte reale ed un'approssimazione della fase minima fornisce risultati che sono molto utili in certi casi pratici (v. cap. 10).

7.4 TRASFORMATE DI HILBERT PER SEQUENZE COMPLESSE

Finora abbiamo considerato le trasformate di Hilbert per la trasformata di Fourier di sequenze causali e per la trasformata di Fourier discreta di sequenze periodiche che sono « causali » nel senso che sono nulle nella seconda metà di ogni periodo. In questo paragrafo esaminiamo le sequenze complesse, le cui componenti reale e immaginaria possono essere legate mediante una convoluzione simile alle trasformate di Hilbert derivate nei paragrafi precedenti. Queste relazioni basate sulla trasformata di Hilbert sono particolarmente utili per la rappresentazione di segnali passa-banda come segnali complessi, in modo del tutto analogo ai « segnali analitici » della teoria dei segnali analogici [6].

Come nelle discussioni precedenti, è possibile far partire la derivazione delle relazioni basate sulla trasformata di Hilbert dal concetto di causalità.

Poiché ci interessa legare le parti reale e immaginaria di una sequenza complessa, la « causalità » sarà applicata alla trasformata di Fourier della sequenza. Non possiamo, naturalmente, imporre che la trasformata di Fourier sia nulla per $\omega < 0$, essendo questa periodica. Definiremo tuttavia la « causalità » in questo contesto intendendo che la trasformata di Fourier vale zero nella seconda metà di ogni periodo, cioè la trasformata z è nulla nella metà inferiore ($-\pi \leq \omega < 0$) del circolo unitario. Perciò, se $s(n)$ indica la sequenza e $S(e^{j\omega})$ la sua trasformata di Fourier, imponiamo che sia

$$S(e^{j\omega}) \equiv 0, \quad -\pi \leq \omega < 0 \quad (7.41)$$

È chiaro che la sequenza $s(n)$ corrispondente a $S(e^{j\omega})$ deve essere complessa in quanto, per la realtà di $s(n)$, occorre che sia $S(e^{-j\omega}) = S^*(e^{j\omega})$. Espriamiamo quindi $s(n)$ come

$$s(n) = s_r(n) + js_i(n) \quad (7.42)$$

dove $s_r(n)$ e $s_i(n)$ sono sequenze reali.

Nella teoria dei segnali analogici il segnale paragonabile a questo è una funzione analitica ed è chiamato perciò *segnale analitico*. In maniera simile, useremo la stessa terminologia per sequenze complesse come $s(n)$. Anche se l'analiticità non ha significato per le sequenze, notiamo che ad ogni sequenza $s(n)$ corrisponde un segnale analogico a banda limitata $s_a(t)$ tale che

$$s_a(t)|_{t=n} = s(n)$$

Allora, se vale che

$$S_a(j\omega) = \begin{cases} S(e^{j\omega}), & 0 \leq \omega < \pi \\ 0, & \text{altrove} \end{cases}$$

il segnale $s_a(t)$ è una funzione analitica di t . In questo senso la sequenza $s(n)$ corrisponde davvero ad un segnale analitico.

Se $S_r(e^{j\omega})$ e $S_i(e^{j\omega})$ indicano le trasformate di Fourier delle sequenze reali $s_r(n)$ e $s_i(n)$, si può facilmente dimostrare che risulta

$$S_r(e^{j\omega}) = \frac{1}{2}[S(e^{j\omega}) + S^*(e^{-j\omega})] \quad (7.43a)$$

e

$$jS_i(e^{j\omega}) = \frac{1}{2}[S(e^{j\omega}) - S^*(e^{-j\omega})] \quad (7.43b)$$

Le trasformate complesse $S_r(e^{j\omega})$ e $S_i(e^{j\omega})$ giocano un ruolo simile a quello che avevano, nei paragrafi precedenti, le parti pari e dispari, rispettivamente, delle sequenze causali. Si noti però che $S_r(e^{j\omega})$ non è una funzione pari ma coniugata pari, cioè $S_r(e^{j\omega}) = S_r^*(e^{-j\omega})$. Analogamente, $jS_i(e^{j\omega})$ è coniugato dispari, cioè $jS_i(e^{j\omega}) = -jS_i^*(e^{-j\omega})$.

Se $S(e^{j\omega})$ vale zero per $-\pi \leq \omega < 0$, allora non vi è sovrapposizione tra le parti non nulle di $S(e^{j\omega})$ e di $S^*(e^{-j\omega})$. Perciò $S(e^{j\omega})$ può essere ricostruito da $S_r(e^{j\omega})$ o da $S_i(e^{j\omega})$. Si noti che poiché $S(e^{j\omega})$ è assunto nullo per $\omega = -\pi$, $S(e^{j\omega})$ si può ricostruire completamente a partire da $jS_i(e^{j\omega})$.

Questo fatto contrasta in qualche misura con le due situazioni precedenti, in cui la funzione causale poteva essere riottenuta dalla sua parte dispari con l'eccezione dei punti estremi.

In particolare risulta

$$S(e^{j\omega}) = \begin{cases} 2S_r(e^{j\omega}), & 0 \leq \omega < \pi \\ 0, & -\pi \leq \omega < 0 \end{cases}$$

e

$$S(e^{j\omega}) = \begin{cases} 2jS_i(e^{j\omega}), & 0 \leq \omega < \pi \\ 0, & -\pi \leq \omega < 0 \end{cases}$$

In alternativa, possiamo legare direttamente $S_r(e^{j\omega})$ e $S_i(e^{j\omega})$, cioè scrivere

$$S_i(e^{j\omega}) = \begin{cases} -jS_r(e^{j\omega}), & 0 \leq \omega < \pi \\ jS_r(e^{j\omega}), & -\pi \leq \omega < 0 \end{cases} \quad (7.44)$$

o

$$S_i(e^{j\omega}) = H(e^{j\omega})S_r(e^{j\omega}) \quad (7.45)$$

dove si è posto

$$H(e^{j\omega}) = \begin{cases} -j, & 0 \leq \omega < \pi \\ j, & -\pi \leq \omega < 0 \end{cases} \quad (7.46)$$

Poiché $S_i(e^{j\omega})$ è la trasformata di Fourier di $s_i(n)$, la parte immaginaria di $s(n)$, e $S_r(e^{j\omega})$ è la trasformata di Fourier di $s_r(n)$, la parte reale di $s(n)$, allora, in base alla (7.45), $s_i(n)$ può essere ottenuto elaborando $s_r(n)$ con un sistema discreto avente risposta in frequenza $H(e^{j\omega})$ fornita dalla (7.46). Questa risposta in frequenza ha modulo unitario, un angolo di fase di $-\pi/2$ per ω tra 0 e π , ed un angolo di fase di $+\pi/2$ per ω tra 0 e $-\pi$. Tale sistema viene spesso chiamato *sfasatore di 90 gradi* o *filtro di Hilbert*. Dalla relazione (7.45) segue che

$$S_r(e^{j\omega}) = \frac{1}{H(e^{j\omega})} S_i(e^{j\omega}) = -H(e^{j\omega})S_i(e^{j\omega}) \quad (7.47)$$

Quindi $-s_r(n)$ può anche essere ottenuto da $s_i(n)$ con uno sfasatore di 90 gradi.

La risposta all'impulso $h(n)$ di uno sfasatore di 90 gradi, corrispondente alla risposta in frequenza definita nella (7.46), è

$$\begin{aligned} h(n) &= \frac{1}{2\pi} \int_{-\pi}^0 j e^{j\omega n} d\omega - \frac{1}{2\pi} \int_0^{\pi} j e^{j\omega n} d\omega \\ &= \frac{2}{\pi} \frac{\sin^2(\pi n/2)}{n}, \quad n \neq 0 \\ &= 0, \quad n = 0 \end{aligned} \quad (7.48)$$

La risposta all'impulso normalizzata è riportata in fig. 7.10. Usando le relazioni (7.45) e (7.47) si ottengono le espressioni

$$s_i(n) = \sum_{m=-\infty}^{\infty} s_r(n-m)h(m) \quad (7.49)$$

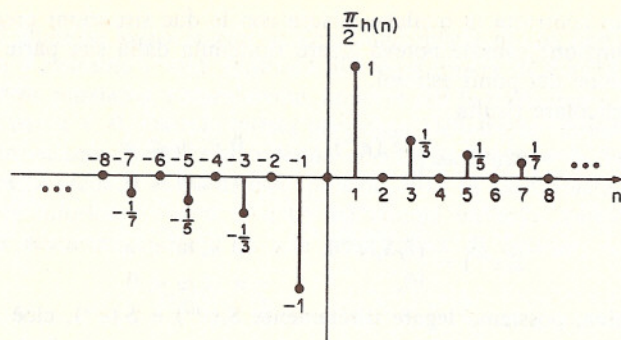


Fig. 7.10 Risposta all'impulso normalizzata di un filtro di Hilbert ideale o sfasatore di 90 gradi.

e

$$s_r(n) = - \sum_{m=-\infty}^{\infty} s_i(n-m)h(m) \quad (7.50)$$

Le (7.49) e (7.50) costituiscono le trasformate di Hilbert per le parti reale e immaginaria di un segnale analitico a tempo discreto.

Una rappresentazione alternativa di $s(n)$ è quella in termini del suo modulo e della sua fase, cioè

$$s(n) = A(n)e^{j\phi(n)} \quad (7.51)$$

dove si è definito

$$A(n) = (s_r^2(n) + s_i^2(n))^{1/2} \quad (7.52a)$$

e

$$\phi(n) = \arctan \left[\frac{s_i(n)}{s_r(n)} \right] \quad (7.52b)$$

La sequenza modulo, $A(n)$, è spesso chiamata *l'involuppo* della sequenza $s(n)$. Il concetto di fase minima discusso nel par. 7.2 ha il suo corrispondente nella teoria dei segnali analitici. Lo sviluppo di questo concetto richiede un formalismo matematico complicato e, poiché non è importante ai fini di questo libro, non lo approfondiremo ulteriormente. Questo argomento è trattato in dettaglio per i segnali a banda limitata da Voelcker [15] ed anche da Requicha [16].

7.4.1 Progetto dei filtri di Hilbert

Si può vedere dall'espressione (7.48) che la trasformata z di $h(n)$ converge *solo* sul circolo unitario. Infatti, a causa della discontinuità della parte immaginaria, la serie

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n}$$

converge alla funzione (7.46) solo in senso quadratico medio. Perciò il filtro di Hilbert ideale o sfasatore di 90 gradi va collocato sullo stesso piano del filtro passa-basso ideale e del derivatore ideale a banda limitata: questi operatori sono importanti dal punto di vista teorico e corrispondono a sistemi non causali per i quali la funzione di trasferimento esiste solo in un senso particolare.

Si possono ovviamente ottenere delle approssimazioni del filtro di Hilbert ideale. Nel caso di approssimazioni di durata finita delle caratte-

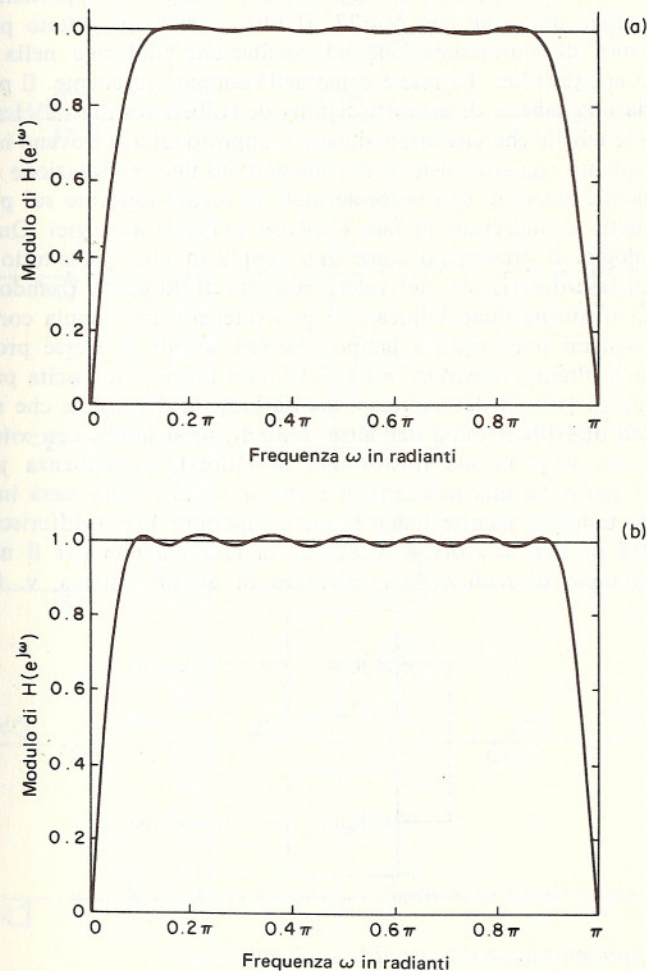


Fig. 7.11 Approssimazioni con risposta all'impulso di durata finita del filtro di Hilbert: (a) $N=27$, progetto con il metodo delle finestre (Blackman); (b) $N=27$, metodo dell'approssimazione con oscillazione uniforme (da Hermann [17]). In entrambi i casi non vi è errore di fase.

ristiche ideali della (7.46), si possono applicare le tecniche standard basate sull'uso di finestre, sul campionamento in frequenza e sull'approssimazione con oscillazione uniforme.

La fig. 7.11(a) mostra un esempio di un filtro di Hilbert progettato col metodo delle finestre, applicando una finestra di Blackman con $N = 27$ alla risposta (7.48) (v. cap. 5). La figura riporta il modulo della risposta in frequenza per $0 \leq \omega \leq \pi$. La fase vale -90° per $0 \leq \omega \leq \pi$ e $+90^\circ$ per $-\pi \leq \omega < 0$. Inoltre, vi è uno sfasamento lineare corrispondente a un ritardo di 13 campioni. La fig. 7.11(b) mostra un'approssimazione ad oscillazione uniforme con $N=27$. Il filtro è stato progettato per avere un errore di approssimazione ad oscillazione uniforme nella banda $0.0874\pi \leq \omega \leq 0.9126\pi$. La fase è come nell'esempio precedente. Il progetto è preso da una tabella di progetti di filtri di Hilbert fornita da Herrmann [17]. Sia le tabelle che una discussione più approfondita si trovano in [18].

Per quanto riguarda sistemi che ammettono una realizzazione ricorsiva, è possibile sfruttare una notevole mole di lavoro esistente sul progetto dei cosiddetti « suddivisori di fase » (*phase splitters*) analogici. Questi sistemi analogici si presentano come una coppia di filtri passa-tutto le cui risposte di fase differiscono del valore costante di 90 gradi. Usando il metodo della trasformazione bilineare si può ottenere una coppia corrispondente di sistemi passa-tutto a tempo discreto, avente le stesse proprietà. Il sistema risultante, illustrato in fig. 7.12, non fornisce un'uscita pari alla trasformata di Hilbert dell'ingresso, ma dà luogo a due uscite che sono la trasformata di Hilbert l'una dell'altra. Quindi, se si indica con $x_r(n)$ l'ingresso e con $x_i(n)$ la sua trasformata di Hilbert, la sequenza $y(n) = y_r(n) + jy_i(n)$ ha una trasformata Z che si annulla sulla metà inferiore del circolo unitario, mentre lungo la metà superiore $Y(e^{j\omega})$ differisce dalla trasformata di $x(n) = x_r(n) + jx_i(n)$ per la fase ma non per il modulo. Per un esempio di realizzazione ricorsiva di questo sistema, v. Gold e altri [8].

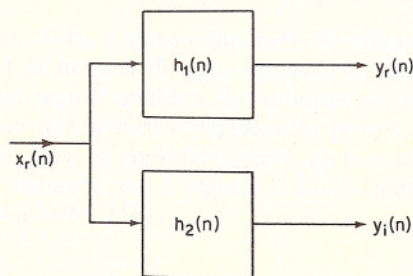


Fig. 7.12 Rappresentazione di un sistema « suddivisore di fase » a 90 gradi

7.4.2 Rappresentazione di segnali passa-banda

Molte applicazioni dei segnali analitici riguardano i segnali a banda stretta impiegati nelle comunicazioni. A volte, in queste applicazioni conviene rappresentare un segnale passa-banda in termini di un segnale passa-

basso. Per vedere come si ottiene tale rappresentazione, consideriamo il segnale complesso passa-basso

$$x(n) = x_r(n) + jx_i(n)$$

dove $x_i(n)$ è la trasformata di Hilbert di $x_r(n)$, e perciò

$$-X(e^{j\omega}) = 0, \quad -\pi \leq \omega < 0$$

Le trasformate di Fourier $X_r(e^{j\omega})$ e $jX_i(e^{j\omega})$ sono rappresentate rispettivamente nelle fig. 7.13(a) e (b), e la trasformata complessiva $X(e^{j\omega}) = X_r(e^{j\omega}) + jX_i(e^{j\omega})$ nella fig. 7.13(c) (le linee a tratto intero rappresen-

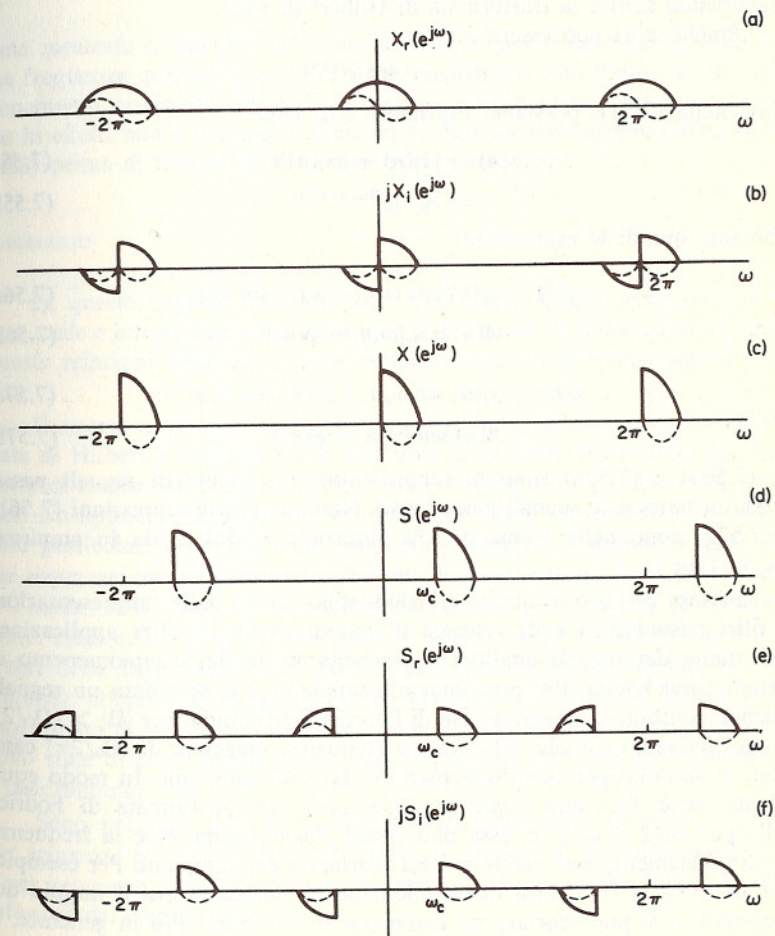


Fig. 7.13 Trasformate di Fourier per la rappresentazione di segnali passa-banda. Le linee a tratto intero indicano le parti reali e quelle tratteggiate le parti immaginarie (si noti che in (b) ed (f) sono riportate le funzioni $jX_i(e^{j\omega})$ e $jS_i(e^{j\omega})$, con $X_i(e^{j\omega})$ e $S_i(e^{j\omega})$ trasformate di Fourier delle trasformate di Hilbert di $x_r(n)$ e $s_r(n)$ rispettivamente)

tano le parti reali e quelle tratteggiate le parti immaginarie). Consideriamo adesso la sequenza (passa-banda)

$$s(n) = x(n)e^{j\omega_c n} = s_r(n) + js_i(n) \quad (7.53)$$

dove $s_r(n)$ e $s_i(n)$ sono sequenze reali. La trasformata di Fourier corrispondente è

$$S(e^{j\omega}) = X(e^{j(\omega - \omega_c)}) \quad (7.54)$$

ed è rappresentata in fig. 7.13(d). Le trasformate di Fourier $S_r(e^{j\omega})$ e $jS_i(e^{j\omega})$ sono mostrate nelle fig. 7.13(e) ed (f). È chiaro che per segnali passa-banda $s_i(n)$ è la trasformata di Hilbert di $s_r(n)$.

Poiché $x(n)$ può essere scritto

$$x(n) = A(n)e^{j\phi(n)}$$

come nella (7.51), possiamo esprimere $s(n)$ come

$$s(n) = [x_r(n) + jx_i(n)]e^{j\omega_c n} \quad (7.55a)$$

$$= A(n)e^{j(\omega_c n + \phi(n))} \quad (7.55b)$$

Abbiamo quindi le espressioni

$$s_r(n) = x_r(n) \cos \omega_c n - x_i(n) \sin \omega_c n \quad (7.56a)$$

$$= A(n) \cos [\omega_c n + \phi(n)] \quad (7.56b)$$

e

$$s_i(n) = x_r(n) \sin \omega_c n + x_i(n) \cos \omega_c n \quad (7.57a)$$

$$= A(n) \sin [\omega_c n + \phi(n)] \quad (7.57b)$$

Le (7.56a) e (7.57a) sono le rappresentazioni cercate di segnali passa-banda in termini di segnali passa-basso. Notiamo che le espressioni (7.56b) e (7.57b) sono nella forma di una sinusoide modulata sia in ampiezza che in fase.

Esempi dell'uso di queste relazioni sono forniti dalle rappresentazioni di filtri passa-banda e di processi di modulazione. Un'altra applicazione importante dei segnali analitici si ha nella teoria del campionamento di segnali passa-banda. Più precisamente, sappiamo che se è dato un segnale a tempo continuo con trasformata di Fourier $X_a(j\Omega)$ nulla per $|\Omega| > (\Omega_0/2)$, allora dobbiamo campionarlo ad una frequenza maggiore di $(\Omega_0/2\pi)$ campioni al secondo per poterlo ricostruire dai suoi campioni. In modo equivalente, se è data una sequenza reale $x_r(n)$ con trasformata di Fourier nulla per $\omega_0/2 \leq \omega \leq \pi$, essa può essere ricampionata, cioè la frequenza di campionamento può essere ridotta scartando dei campioni. Per esempio, se è $\omega_0 = \pi$, la frequenza di campionamento originaria era il doppio del necessario e si può scartare un campione su due. Più in generale, il numero di campioni al secondo può essere ridotto di un fattore $2\pi/\omega_0$.

Consideriamo adesso un segnale passa-banda reale $s_r(n)$ come mostrato in fig. 7.13(e). Essendo il segnale reale, la sua trasformata di Fourier è ovviamente coniugata simmetrica. Per determinare la frequenza di cam-

pionamento minima, dobbiamo considerare come larghezza di banda dello spettro la quantità $\omega_0 = 2(\omega_c + \Delta\omega)$; cioè in questo caso, anche se la larghezza di banda effettiva è $\Delta\omega$, si può ridurre la frequenza di campionamento di $s_r(n)$ solo del fattore $2\pi/\omega_0$. Consideriamo invece il segnale analitico $s(n) = s_r(n) + js_i(n)$ con trasformata di Fourier $S(e^{j\omega})$ come mostrato in fig. 7.13(d). Poiché $S(e^{j\omega})$ è nulla ovunque eccetto che nella regione $\omega_c \leq \omega \leq \omega_c + \Delta\omega$, la frequenza di campionamento può essere ridotta di un fattore $2\pi/\Delta\omega$. Si può vedere questo notando che il segnale analitico complesso passa-basso è

$$x(n) = s(n)e^{-j\omega_c n} \quad (7.58)$$

come mostrato in fig. 7.13(c). Questo segnale può essere campionato ad una frequenza minore di un fattore $2\pi/\Delta\omega$ rispetto alla frequenza di campionamento originaria. Facendo qualche ulteriore considerazione si vede che in effetti non è necessario eseguire la modulazione rappresentata dalla (7.58) prima di ridurre la frequenza di campionamento.

SOMMARIO

In questo capitolo abbiamo preso in esame diverse relazioni tra le parti reale e immaginaria di trasformate di Fourier e di sequenze complesse. Queste relazioni vengono indicate sinteticamente come *trasformate di Hilbert*.

Il nostro approccio alla derivazione di tutte le forme della trasformata di Hilbert è stato quello di utilizzare la nozione di causalità che permette di ricostruire una sequenza o una funzione a partire dalla sua parte pari. Un approccio alternativo, sviluppato in alcuni dei problemi, è basato sulle particolari proprietà delle funzioni analitiche. Abbiamo trovato che per sequenze causali la parte reale e quella immaginaria della trasformata di Fourier sono legate da una relazione simile a un integrale di convoluzione. Inoltre, per il caso particolare di sequenze sia causali che con poli e zeri della trasformata z interni al cerchio unitario (condizione di fase minima), abbiamo dimostrato che il logaritmo del modulo e la fase della trasformata di Fourier costituiscono una coppia di trasformate di Hilbert. Sono state discusse anche molte altre importanti proprietà delle sequenze a fase minima.

Sono poi state derivate la forma delle trasformate di Hilbert valide per sequenze periodiche che soddisfano un vincolo di causalità modificato e per sequenze complesse con trasformata di Fourier nulla lungo la metà inferiore del cerchio unitario.

Nel discutere le trasformate di Hilbert ci siamo soffermati più sui concetti teorici che sulle applicazioni. Alcuni usi delle trasformate di Hilbert sono illustrati nei problemi e anche nel cap. 10, dove i risultati di questo capitolo hanno un ruolo molto importante.

BIBLIOGRAFIA

1. L. V. Ahlfors, *Complex Analysis*, 2nd ed., McGraw-Hill Book Company, New York, 1966.
2. R. V. Churchill, *Complex Variables and Applications*, McGraw-Hill Book Company, New York, 1960.
3. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill Book Company, New York, 1953.
4. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand Reinhold Company, New York, 1945.
5. E. A. Guillemin, *Theory of Linear Physical Systems*, John Wiley & Sons, Inc., New York, 1963.
6. J. Dugundji, "Envelopes and Pre-Envelopes of Real Waveforms," *Trans. IRE*, Vol. IT-4, Mar. 1958, pp. 53-57.
7. V. Cizek, "Discrete Hilbert Transform," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, No. 4, Dec. 1970, pp. 340-343.
8. B. Gold, A. V. Oppenheim, and C. M. Rader, "Theory and Implementation of the Discrete Hilbert Transform," *Proc. Symp. Computer Processing in Communications*, Vol. 19, Polytechnic Press, 1970, New York.
9. D. J. Sakrison, W. T. Ford, and J. H. Hearne, "The z-Transform of a Realizable Time Function," *IEEE Trans. Geosci. Elect.*, Vol. GE-5, No. 2, Sept. 1967, pp. 33-41.
10. F. B. Hildebrand, *Advanced Calculus with Applications*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.
11. S. Treitel and E. A. Robinson, "The Design of High-Resolution Digital Filters," *IEEE Trans. Geosci. Elect.*, Vol. GE-4, No. 1, June 1966, pp. 25-38.
12. E. A. Robinson, *Random Wavelets and Cybernetic Systems*, Charles Griffin and Co. Ltd., London, 1962.
13. E. A. Robinson, *Statistical Communication and Detection*, Hafner Press, New York, 1967.
14. A. J. Berkhout, "On the Minimum Phase Criterion of Sampled Signals," *IEEE Trans. Geosci. Elect.*, Vol. GE-11, No. 4, Oct. 1973, pp. 186-198.
15. H. B. Voelcker, "Toward a Unified Theory of Modulation," *Proc. IEEE*, Vol. 54, Mar. 1966, pp. 340-353, and May 1966, pp. 735-755.
16. A. A. G. Requicha, "Contributions to a Zero-Based Theory of Bandlimited Signals," Ph.D. Thesis, Department of Electrical Engineering, University of Rochester, 1970.
17. O. Herrmann, "Transversalfilter zur Hilbert-Transformation," *Arch. Elektronik Übertragungstechnik*, Vol. 23, No. 12, 1969, pp. 581-587.
18. L. R. Rabiner and R. W. Schafer, "On the Behavior of Minimax FIR Digital Hilbert Transformers," *Bell Syst. Tech. J.*, Vol. 53, No. 2, Febr., 1974, pp. 361-388.
19. O. Herrmann and H. W. Schuessler, "Design of Nonrecursive Digital Filters with Minimum Phase," *Elect. Letters*, Vol. 6, No. 11, 1970, pp. 329-330.

PROBLEMI

1. Nel par. 7.1 abbiamo osservato che la trasformata z è completamente determinata fuori del cerchio unitario dal valore della sua parte immaginaria sul circolo unitario e dal valore $h(0)$.

- (a) Derivare la relazione (7.14b) partendo dalla (7.8).
- (b) Usare la relazione (7.14b) per trovare $H(z)$ quando è

$$H_I(e^{j\omega}) = \frac{-\alpha \sin \omega}{1 + \alpha^2 - 2\alpha \cos \omega}$$

$$h(0) = 1$$

2. Nel par. 7.1 abbiamo ricavato la trasformata di Hilbert di una sequenza $h(n)$ reale e causale usando la proprietà che le parti reale e immaginaria di $H(z)$ sul circolo unitario corrispondono, rispettivamente, alle parti pari e dispari di $h(n)$. In alternativa, essendo la trasformata z analitica nella sua regione di convergenza, possiamo ricavare direttamente la trasformata di Hilbert nel dominio della frequenza, usando la formula integrale di Cauchy. La formula integrale di Cauchy dice che se $F(z)$ è analitica dovunque dentro e lungo un percorso chiuso C , allora risulta

$$\frac{1}{2\pi j} \oint_C \frac{F(\zeta) d\zeta}{\zeta - z} = \begin{cases} F(z), & \text{se } z \text{ è interno a } C \\ 0, & \text{se } z \text{ è esterno a } C \end{cases}$$

Sia $H(z)$ la trasformata di una sequenza stabile, causale e complessa $h(n)$.

- (a) Poniamo $F(z) = H(1/z)$. $F(z)$ è analitica in tutta la regione $|z| < R$, con $R > 1$. Perché?
- (b) Sia il contorno C il circolo unitario $|z| = 1$. Se il punto ζ appartiene a C , dimostrare che il punto

$$\bar{z} = \frac{\zeta \zeta^*}{z^*}$$

cade all'esterno di C per ogni z interno a C . I punti z e \bar{z} si dicono punti inversi (o punti immagine) rispetto alla circonferenza C .

- (c) Considerando l'espressione

$$\frac{1}{2\pi j} \oint_C \frac{F(\zeta) d\zeta}{\zeta - z} + \frac{\alpha}{2\pi j} \oint_C \frac{F(\zeta) d\zeta}{\zeta - \bar{z}}$$

per $\alpha = +1$ e $\alpha = -1$, dimostrare che si può trovare $F(z)$ ovunque dentro il circolo unitario a partire da $F(e^{j\theta})$ usando le convoluzioni

$$F(re^{j\theta}) = \begin{cases} \frac{1}{2\pi} \int_{-\pi}^{\pi} P(r, \theta - \phi) F(e^{j\phi}) d\phi, & r < 1 \quad (\text{P7.2-1}) \\ F(0) + \frac{j}{2\pi} \int_{-\pi}^{\pi} Q(r, \theta - \phi) F(e^{j\phi}) d\phi, & r < 1 \quad (\text{P7.2-2}) \end{cases}$$

dove è

$$\begin{aligned} z &= re^{j\theta}, & 0 \leq r < 1, & -\pi < \theta \leq \pi \\ \zeta &= e^{j\phi}, & -\pi < \phi \leq \pi \end{aligned}$$

e P e Q sono reali. Trovare $P(r, \theta)$ e $Q(r, \theta)$.

Suggerimento:

$$\frac{1}{\zeta - z} + \frac{\alpha}{\zeta - \bar{z}} = \frac{1}{\zeta} \frac{\zeta}{\zeta - z} - \alpha \frac{\zeta^*}{\zeta^* - z^*} + \alpha$$

(d) Poniamo $F(re^{j\theta}) = u(r, \theta) + jv(r, \theta)$. Quando è $r < 1$, usare le relazioni (P7.2-1) e (P7.2-2) per trovare le espressioni di

- (1) $u(r, \theta)$ in termini di $u(1, \theta)$.
- (2) $v(r, \theta)$ in termini di $v(1, \theta)$.
- (3) $u(r, \theta)$ in termini di $v(1, \theta)$ e $u(0)$.
- (4) $v(r, \theta)$ in termini di $u(1, \theta)$ e $v(0)$.
- (5) $F(re^{j\theta})$ in termini di $u(1, \theta)$ e $v(0)$.
- (6) $F(re^{j\theta})$ in termini di $v(1, \theta)$ e $u(0)$.

Potete esprimere le risposte alle parti (e) ed (f) usando la relazione

$$K(re^{j\theta}) = P(r, \theta) + jQ(r, \theta)$$

(e) Sia $H(\rho e^{j\omega}) = H_R(\rho, \omega) + jH_I(\rho, \omega)$. Usando il fatto che è $H(\rho e^{j\omega}) = F(\rho^{-1}e^{-j\omega})$, trasformate i risultati relativi alla parte (d) per ottenere le trasformate di Hilbert sul cerchio unitario.

[Si noti che è $K(z^{-1}) = U(z)$, dove $U(z)$ è il nucleo definito nel testo].

(f) Dovreste aver trovato che le costanti $H_R(\infty)$ e $H_I(\infty)$ (o le loro equivalenti nel dominio del tempo) compaiono in alcune relazioni della parte (e). Dimostrate che se $h(n)$ è reale, $H_I(\infty) = 0$ e se $h(n)$ è immaginario $H_R(\infty) = 0$. (Suggerimento: usate il teorema del valore iniziale, se non lo avete ancora fatto).

3. Supponiamo che $F(z)$ sia una funzione razionale, cioè

$$F(z) = \frac{N(z)}{D(z)}$$

dove $N(z)$ e $D(z)$ sono polinomi. Assumiamo inoltre che $F(z)$ non abbia poli o zeri di molteplicità superiore a 1. Sia C una curva chiusa, e siano Z e P il numero, rispettivamente, di zeri e poli di $F(z)$ compresi dentro la curva (assumiamo che non vi siano poli o zeri su C).

(a) Dimostrare che risulta

$$\frac{1}{2\pi j} \oint_C \frac{F'(z)}{F(z)} dz = Z - P$$

(b) Se $F(z) = |F(z)|e^{j\arg F(z)}$, dimostrare che è

$$\frac{1}{2\pi} \arg [F(z)] \Big|_C = Z - P$$

cioè la variazione di fase di F quando la curva C è percorsa esattamente una volta vale $2\pi(Z - P)$.

(Si può dimostrare che questi risultati si generalizzano al caso di poli o zeri multipli contando i poli e gli zeri in base alla loro molteplicità: ad es., un polo del secondo ordine è contato due volte).

4. Sia $x(n)$ una sequenza reale causale per cui $|x(n)| < \infty$. La trasformata z di $x(n)$ è

$$X(z) = \sum_{n=0}^{\infty} x(n)z^{-n}$$

che è una serie di Taylor nella variabile z^{-1} , e perciò converge ad una funzione analitica ovunque all'esterno di una corona circolare centrata in $z = 0$ [la regione di convergenza comprende il punto $z = \infty$ ed in effetti è $X(\infty) = x(0)$]. L'affermazione che $X(z)$ è analitica (nella regione di convergenza) implica dei forti vincoli su X : le sue parti reale ed immaginaria soddisfanno ciascuna l'equazione di Laplace e sono legate tra loro dalle equazioni di Cauchy-Riemann. Useremo ora queste proprietà per ricavare $X(z)$ dalla sua parte reale quando $x(n)$ è una sequenza reale, causale a valori finiti.

Sia $x(n)$ una sequenza reale (a valori finiti) causale con trasformata z

$$X(z) = X_R(z) + jX_I(z)$$

dove X_R e X_I sono funzioni di z a valori reali.

Supponiamo che $X_R(z)$ sia espressa da

$$X_R(\rho e^{j\omega}) = \frac{\rho + \alpha \cos \omega}{\rho} \quad (\alpha \text{ reale})$$

per $z = \rho e^{j\omega}$. Trovare $X(z)$ (come funzione esplicita di z) assumendo che sia analitica ovunque eccetto che in $z = 0$ e usando entrambi i metodi suggeriti nel seguito.

(a) *Metodo 1* (nel dominio della frequenza): usare il fatto che le parti reale e immaginaria di X devono soddisfare le equazioni di Cauchy-Riemann in tutta la regione di analiticità di X . Le equazioni di Cauchy-Riemann sono le seguenti:

(1) in coordinate cartesiane:

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y}$$

$$\frac{\partial V}{\partial x} = -\frac{\partial U}{\partial y}$$

dove è $z = x + jy$ e $X(x + jy) = U(x, y) + jV(x, y)$.

(2) in coordinate polari:

$$\frac{\partial U}{\partial \rho} = \frac{1}{\rho} \frac{\partial V}{\partial \omega}$$

$$\frac{\partial V}{\partial \rho} = -\frac{1}{\rho} \frac{\partial U}{\partial \omega}$$

dove è $z = \rho e^{j\omega}$ e $X(\rho e^{j\omega}) = U(\rho, \omega) + jV(\rho, \omega)$.

Poiché sappiamo che è $U = X_R$, possiamo integrare queste equazioni per ricavare V e quindi X . (Attenzione a considerare correttamente la costante di integrazione).

(b) *Metodo 2* (nel dominio del tempo): usare il fatto che la sequenza $x_e(n)$, con trasformata di Fourier $X_R(e^{j\omega})$, deve essere reale e pari e che la sequenza $x_o(n)$, con trasformata di Fourier $jX_I(e^{j\omega})$, è reale e dispari. Per la linearità risulta

$$x(n) = x_e(n) + x_o(n)$$

Poiché possiamo ricavare $x_e(n)$ direttamente da $X_R(e^{j\omega})$ e poiché $x(n)$ è reale e causale, possiamo trovare $x_o(n)$ e quindi $X(z)$.

5. Ricavate un'espressione integrale per $H(z)$ all'interno del cerchio unitario in termini di $\text{Re}[H(e^{j\omega})]$, quando $h(n)$ è una sequenza reale e stabile tale che $h(n) = 0$ per $n > 0$.

6. Dimostrare che la funzione di trasferimento passa-tutto

$$H_{ap}(z) = \frac{z^{-1} - a^*}{1 - az^{-1}}, \quad |a| < 1$$

ha guadagno unitario a tutte le frequenze. In altri termini, dimostrare che $|H_{ap}(e^{j\omega})| = 1$ per $0 \leq \omega \leq \pi$.

7. Si consideri un segnale $x(n)$ causale, stabile e non a fase minima, con trasformata z data da $X(z)$. Gli zeri di $X(z)$ sono z_k , $k = 1, 2, \dots, M$, con $|z_1| < |z_2| < \dots < |z_M|$. Si vuole ottenere una nuova sequenza $y(n)$ che sia a fase minima pesando esponenzialmente la sequenza $x(n)$:

$$y(n) = \alpha^n x(n)$$

Come si deve scegliere α in modo che $y(n)$ sia a fase minima?

8. Nel par. 7.2 è stato detto che le sequenze a fase minima si chiamano così in quanto le loro funzioni di ritardo di fase (l'opposto della fase) sono le più piccole rispetto a tutte le sequenze reali e causali aventi trasformata di Fourier con uguale modulo. Questa affermazione poggia sulla proprietà che la fase di un sistema passa-tutto reale, causale e stabile è sempre minore o uguale a zero.
- (a) Si consideri la funzione di trasferimento passa-tutto del primo ordine

$$H_{ap}(z) = \frac{z^{-1} - a}{1 - az^{-1}}$$

dove a è reale e $|a| < 1$. Dimostrare che la fase di $H_{ap}(e^{j\omega})$ è non positiva per $0 \leq \omega < \pi$.

- (b) Si consideri ora la funzione di trasferimento passa-tutto del secondo ordine con poli e zeri complessi coniugati

$$H_{ap}(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \frac{z^{-1} - a}{1 - a^*z^{-1}}$$

dove a è complesso e $|a| < 1$. Dimostrare che la fase di $H_{ap}(e^{j\omega})$ è non positiva per $0 \leq \omega < \pi$.

Poiché la funzione di trasferimento di qualunque sistema passa-tutto con risposta all'impulso reale può sempre essere rappresentata come un prodotto di fattori del tipo di quelli delle parti (a) e (b), i risultati di sopra dimostrano che se è

$$H(z) = H_{\min}(z)H_{ap}(z)$$

allora la fase di $H(e^{j\omega})$ sarà sempre più negativa della fase di $H_{\min}(e^{j\omega})$, e perciò $H_{\min}(e^{j\omega})$ ha il ritardo di fase minimo fra tutti i sistemi per cui è $|H(e^{j\omega})| = |H_{\min}(e^{j\omega})|$.

(Suggerimento: sia in (a) che in (b) si può arrivare al risultato desiderato mediante manipolazioni algebriche di $\arg[H(e^{j\omega})]$. In alternativa, è sufficiente una semplice considerazione geometrica dopo aver trasformato le funzioni di trasferimento nel piano s usando la trasformazione bilineare).

9. Dimostrare la validità delle due affermazioni seguenti:
- La convoluzione di due sequenze a fase minima è ancora a fase minima.
 - La somma di due sequenze a fase minima non è necessariamente a fase minima. (Dare un esempio di una sequenza a fase minima e di una a fase non minima ottenibili come somma di due sequenze a fase minima).
10. Sia $h_{\min}(n)$ una sequenza a fase minima con trasformata z data da $H_{\min}(z)$. Se $h(n)$ è una sequenza causale a fase non minima, la cui trasformata di Fourier ha modulo $|H_{\min}(e^{j\omega})|$, dimostrare che risulta

$$|h(0)| < |h_{\min}(0)|$$

(Suggerimento: usare il teorema del valore iniziale).

11. Una delle proprietà interessanti ed importanti delle sequenze a fase minima è quella del minimo ritardo dell'energia: tra le sequenze causali con ugual modulo $|H(e^{j\omega})|$ della trasformata di Fourier, la quantità

$$\sum_{n=0}^m |h(n)|^2$$

è massima quando $h(n)$ è la sequenza a fase minima. Questo risultato si dimostra come segue: sia $h_{\min}(n)$ una sequenza a fase minima con trasformata z data da $H_{\min}(z)$. Inoltre, sia z_k uno zero di $H_{\min}(z)$, in modo che $H_{\min}(z)$ può essere espresso come

$$H_{\min}(z) = Q(z)(1 - z_k z^{-1}), \quad |z_k| < 1$$

dove $Q(z)$ è ancora a fase minima. Consideriamo adesso un'altra sequenza $h(n)$ con trasformata $H(z)$ tale che sia

$$|H(e^{j\omega})| = |H_{\min}(e^{j\omega})|$$

e $H(z)$ abbia uno zero in $z = 1/z_k^*$ invece che in z_k .

- Esprimere $H(z)$ in termini di $Q(z)$.
- Esprimere $h(n)$ ed $h_{\min}(n)$ in termini della sequenza a fase minima $q(n)$ avente $Q(z)$ per trasformata z .
- Per confrontare la distribuzione di energia delle due sequenze, dimostrare che

$$\varepsilon = \sum_{n=0}^m |h_{\min}(n)|^2 - \sum_{n=0}^m |h(n)|^2 = (1 - |z_k|^2) |q(m)|^2$$

- Usando il risultato della parte (c), dimostrare la disuguaglianza

$$\sum_{n=0}^m |h(n)|^2 \leq \sum_{n=0}^m |h_{\min}(n)|^2, \quad \text{per qualsiasi } m$$

12. La fig. P7.12 mostra otto diverse sequenze di durata finita. Ogni sequenza è lunga quattro punti. Il modulo della trasformata di Fourier è lo stesso per tutte le sequenze. Quale sequenza ha tutti gli zeri della sua trasformata z dentro il cerchio unitario?
13. Siano $h_1(n)$, $h_2(n)$, ..., $h_M(n)$ M sequenze di durata finita, tutte lunghe N , cioè con $h_i(n) = 0$ per $n < 0$ e $n \geq N$. Il modulo della trasformata di Fourier è identico per ogni sequenza. Nessuna sequenza è proporzionale (tramite una costante

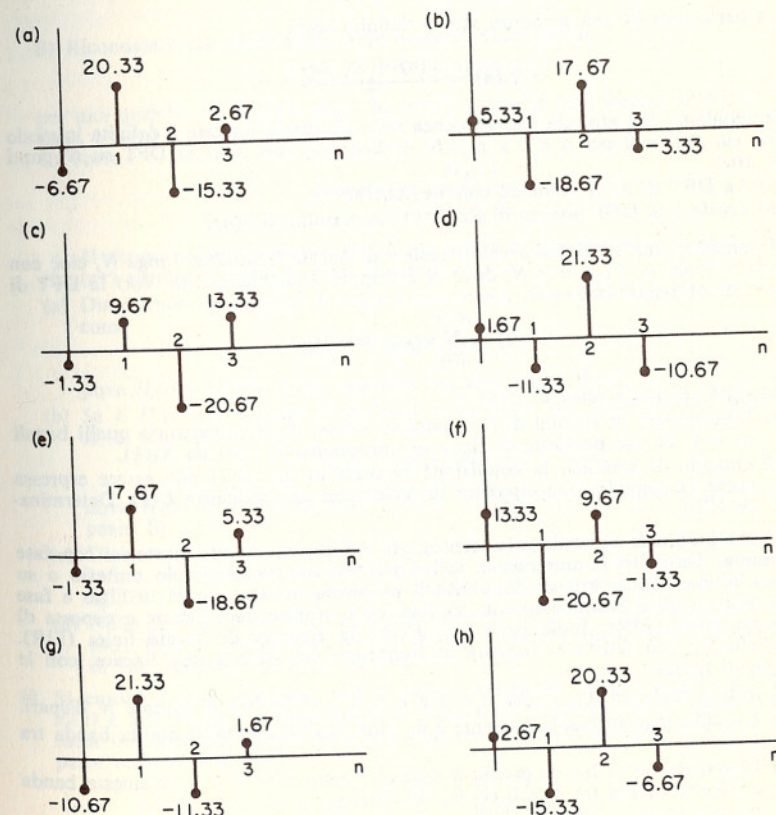


Fig. P7.12

di proporzionalità reale o complessa) a una delle altre.

- (a) Quale è il massimo valore di M se non si impone che le sequenze siano reali?
 (b) Quale è il massimo valore di M se si impone che le sequenze siano reali?

14. Una sequenza a fase massima si ottiene ribaltando tutti gli zeri della trasformata z di una sequenza a fase minima nelle posizioni coniugate reciproche all'esterno del cerchio unitario. In altri termini, possiamo esprimere la trasformata z di una sequenza a fase massima come

$$H_{\max}(z) = H_{\min}(z)H_{\text{ap}}(z)$$

Nel caso di una sequenza di durata finita, $H_{\min}(z)$ è esprimibile come

$$H_{\min}(z) = h_{\min}(0) \prod_{k=1}^{N-1} (1 - z_k z^{-1}), \quad |z_k| < 1$$

- (a) Ricavare un'espressione per la funzione passa-tutto necessaria per ottenere $H_{\max}(z)$.

- (b) Mostrare che $H_{\max}(z)$ può essere scritto come

$$H_{\max}(z) = z^{-(N-1)} H_{\min}(z^{-1})$$

- (c) Usando il risultato della parte (b) esprimere la sequenza a fase massima $h_{\max}(n)$ in termini di $h_{\min}(n)$.

15. La parte pari di una sequenza $x(n)$ è definita come

$$x_e(n) = \frac{x(n) + x(-n)}{2}$$

Supponiamo che $x(n)$ sia una sequenza reale di durata limitata e definita in modo che sia $x(n) = 0$ per $n < 0$ e $n \geq N$. Indichiamo con $X(k)$ la DFT su N punti di $x(n)$.

- (a) La DFT di $x_e(n)$ coincide con $\text{Re}[X(k)]$?

- (b) Quale è la DFT inversa di $\text{Re}[X(k)]$ in termini di $x(n)$?

16. Si consideri una sequenza a valori reali e di durata finita $x(n)$ lunga N , cioè con $x(n) = 0$ per $n < 0$ e $n \geq N$, dove N è dispari. Indichiamo con $X(k)$ la DFT di $x(n)$ su M punti, cioè

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/M)nk}$$

Sia $X_R(k)$ la parte reale di $X(k)$.

- (a) Determinare in termini di N il minimo valore di M (diverso da quelli banali $M = 1, 2$) che permette di ricavare univocamente $X(k)$ da $X_R(k)$.

- (b) Quando M soddisfa la condizione ricavata in (a), $X(k)$ può essere espressa come la convoluzione circolare di $X_R(k)$ con una sequenza $U(k)$. Determinare $U(k)$.

17. Questo problema riguarda una tecnica per il progetto di filtri numerici con fase minima. Tali filtri hanno tutti i poli e gli zeri interni al cerchio unitario o su quest'ultimo. Consideriamo dapprima il problema di trasformare in filtro a fase minima un filtro passa-basso con oscillazione uniforme, fase lineare e risposta di durata finita (FIR). Indichiamo con $H(e^{j\omega})$ la risposta di durata finita (FIR). Indichiamo con $H(e^{j\omega})$ la risposta in frequenza del filtro a fase lineare, con le seguenti ipotesi:

- (1) $h(n)$ è reale. $h(n) = 0$ per $n < 0$ e $n > N - 1$. Inoltre si assume N dispari.
 (2) L'oscillazione in banda passante è δ_1 , cioè $|H(e^{j\omega})|$ oscilla in questa banda tra $1 + \delta_1$ e $1 - \delta_1$.
 (3) L'oscillazione in banda oscura è δ_2 , cioè risulta $|H(e^{j\omega})| < \delta_2$ in questa banda e $|H(e^{j\omega})|$ oscilla tra δ_2 e 0 (v. fig. P7.17-1).
 (4) Risulta $H(e^{j\omega}) = H_0(e^{j\omega})e^{-jn_0\omega}$, dove $H_0(e^{j\omega})$ è reale e $n_0 = (N - 1)/2$.

Herrmann e Schuessler [19] hanno proposto la seguente tecnica per trasfor-

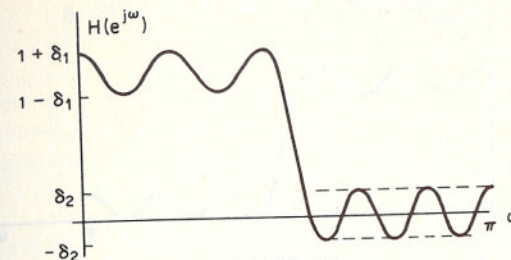


Fig. P7.17-1

mare questo filtro a fase lineare in uno a fase minima con funzione di trasferimento $H_{\min}(z)$ e risposta all'impulso $h_{\min}(n)$.

- A) Generare una nuova sequenza

$$h_1(n) = \begin{cases} h(n), & n \neq n_0 \\ h(n_0) + \delta_2, & n = n_0 \end{cases}$$

- B) Riconoscere che $H_1(z)$ può essere espressa nella forma

$$H_1(z) = z^{-n_0} H_2(z) H_2(1/z)$$

per una qualche $H_2(z)$, dove $H_2(z)$ ha tutti i poli e gli zeri interni al cerchio unitario, e $h_2(n)$ è reale.

- C) Porre

$$H_{\min}(z) = \frac{H_2(z)}{\sqrt{1 + \delta_2}}$$

Il denominatore serve per rinormalizzare la banda passante, in modo che la $H_{\min}(e^{j\omega})$ risultante oscilli intorno al valore uno.

- (a) Dimostrare che se $h_1(n)$ è scelto come in A), allora $H_1(e^{j\omega})$ può essere scritta come

$$H_1(e^{j\omega}) = e^{-jn_0\omega} H_3(e^{j\omega})$$

dove $H_3(e^{j\omega})$ è reale e non negativa per ogni valori di ω .

- (b) Se è $H_3(e^{j\omega}) \geq 0$, come dimostrato nella parte (a), mostrare che esiste una $H_2(z)$ tale che

$$H_3(z) = H_2(z) H_2(1/z)$$

dove $H_2(z)$ è a fase minima e $h_2(n)$ è reale (ciò equivale a giustificare il passo B).

- (c) Dimostrare che il nuovo filtro $H_{\min}(e^{j\omega})$ è un passa-basso con oscillazione uniforme, cioè che la sua caratteristica di ampiezza è della forma illustrata in fig. P7.17-2, calcolando δ_1' e δ_2' . Quale è N' , cioè la lunghezza di $h_{\min}(n)$?
 (d) Nelle parti (a), (b) e (c) abbiamo assunto che il filtro di partenza fosse un filtro FIR a fase lineare. Questa tecnica funziona anche togliendo il vincolo di fase minima? Motivare la risposta.

18. Si consideri una sequenza $x(n)$ a valori complessi: $x(n) = x_r(n) + jx_i(n)$, con $x_r(n)$ e $x_i(n)$ reali. La trasformata z , $X(z)$, della sequenza $x(n)$ è nulla lungo la metà inferiore del cerchio unitario, cioè risulta $X(e^{j\omega}) = 0$ per $\pi \leq \omega \leq 2\pi$. La parte reale di $x(n)$ è

$$x_r(n) = \begin{cases} \frac{1}{2}, & n = 0 \\ -\frac{1}{4}, & n = \pm 2 \\ 0, & \text{altrove} \end{cases}$$

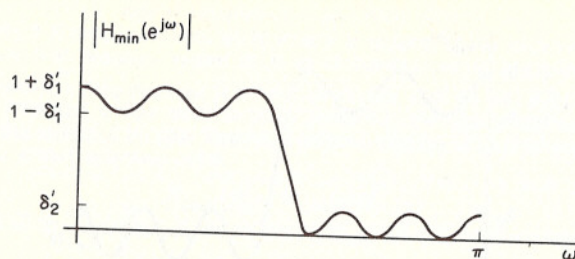


Fig. P7.17-2

- Ricavare le parti reale e immaginaria di $X(e^{j\omega})$.
 19. Indichiamo con $H[\]$ l'operazione ideale della trasformata di Hilbert, cioè

$$H[x(n)] = \sum_{k=-\infty}^{\infty} h(n-k)x(k)$$

dove $h(n)$ è fornita dall'espressione (7.48). Dimostrare le seguenti proprietà.

(a) $H[H[x(n)]] = -x(n)$.

(b) $\sum_{n=-\infty}^{\infty} x(n)H[x(n)] = 0$.

(Suggerimento: usare il teorema di Parseval).

(c) $H[x(n) * y(n)] = H[x(n)] * y(n) = x(n) * H[y(n)]$, dove $x(n)$ e $y(n)$ sono sequenze arbitrarie.

8. SEGNALI CASUALI DISCRETI

8.0 INTRODUZIONE

Nei precedenti capitoli ci siamo occupati delle rappresentazioni matematiche dei segnali e sistemi a tempo discreto, delle loro implicazioni e del loro uso. Abbiamo visto come segnali e sistemi a tempo discreto hanno una rappresentazione sia nel dominio del tempo che in quello della frequenza, e che ciascuna di queste gioca un ruolo importante nella teoria e nel progetto dei sistemi di elaborazione numerica dei segnali. Finora abbiamo assunto che i segnali siano deterministici, vale a dire che ogni valore di una sequenza è univocamente determinato o da una espressione matematica, o da una tavola di dati o per mezzo di una regola di qualche tipo. Nei capitoli 1 e 2 abbiamo in particolare discusso la rappresentazione di tali segnali deterministici in termini delle loro trasformate z o trasformate di Fourier. Affinché una sequenza ammetta una rappresentazione mediante trasformata z occorre che essa abbia energia finita oppure che esista una sequenza esponenziale tale che il prodotto con essa abbia energia finita. Questo vincolo sull'energia corrisponde in sostanza al vincolo che la trasformata z converga. Nel cap. 3 abbiamo poi considerato le sequenze periodiche. La trasformata z di una sequenza periodica non esiste, in quanto la condizione sull'energia finita non può essere soddisfatta. Tuttavia i segnali periodici sono, per definizione, identici da periodo a periodo, e quindi possono essere rappresentati univocamente in termini di un singolo periodo. Un singolo periodo è una sequenza di durata finita e quindi anche di energia finita. Questa proprietà dei segnali periodici è stata utilizzata nel cap. 3 per darne una rappresentazione per mezzo della serie di Fourier o della trasformata di Fourier discreta.

Esistono molti esempi importanti di segnali che o non hanno energia finita o non sono periodici. Molti segnali nel campo delle comunicazioni, per esempio, hanno durata infinita e pertanto si modellano meglio in termini di segnali di energia e durata infinite. In numerose situazioni i processi che danno origine ai segnali sono così complessi da rendere estremamente difficile, se non impossibile, la descrizione precisa di un segnale. Come esempio, vedremo, nel cap. 9, che molti degli effetti di imprecisione cui si va incontro quando si realizzano algoritmi di elaborazione numerica dei segnali con registri a lunghezza finita, possono essere rappresentati mediante « rumore » additivo, per il quale la descrizione più conveniente è

proprio quella di una sequenza a energia infinita. Molti sistemi meccanici generano segnali acustici o di vibrazione che si possono spesso elaborare per diagnosticare guasti potenziali; anche qui, segnali di questo tipo si modellano nel modo più conveniente in termini di segnali non periodici a energia infinita. Due ulteriori esempi fra i numerosi ancora possibili sono il segnale voce, da elaborare a scopo di riconoscimento automatico o compressione di banda, e la musica, da elaborare, per esempio, per migliorarne la qualità.

La chiave per la rappresentazione matematica di questo tipo di segnali sta nella loro descrizione in termini di medie. Come vedremo nel seguito di questo capitolo, molte (ma non tutte) le proprietà di questi segnali si possono esprimere sinteticamente per mezzo di una sequenza a energia finita chiamata sequenza di *autocorrelazione* o di *autocovarianza*, per la quale spesso esiste la trasformata z o la trasformata di Fourier. Come anche vedremo, la trasformata di Fourier della sequenza di autocovarianza ha una interessante interpretazione in termini di distribuzione in frequenza della potenza del segnale. L'uso della sequenza di autocovarianza e della sua trasformata ha l'ulteriore importante vantaggio di consentire di descrivere l'elaborazione operata da un sistema lineare discreto su segnali a energia infinita mediante l'effetto di quel sistema sulla sequenza di autocovarianza dei medesimi segnali.

Nello sviluppare la rappresentazione dei segnali a energia infinita, è conveniente lavorare nell'ambito dei segnali non deterministici, cioè casuali o aleatori. Da questo punto di vista il segnale è considerato come membro di un insieme di segnali a tempo discreto caratterizzato da un insieme di funzioni di densità di probabilità. Nella sua formulazione più generale, la teoria dei segnali casuali è estremamente avanzata e astratta, e una trattazione rigorosa richiederebbe un grado di sofisticazione matematica che non è negli scopi di questo libro. Il principale obiettivo in questo capitolo sarà quindi quello di raccogliere e interpretare uno specifico insieme di risultati che riguardano la rappresentazione dei segnali casuali a energia infinita, risultati che ci saranno utili nei capitoli successivi. Eviteremo pertanto la discussione dettagliata della maggior parte dei difficili e sottili problemi matematici della teoria dei processi casuali. Pur con un approccio non necessariamente rigoroso, si sarà tuttavia alquanto accurati nel riassumere i risultati più importanti e le ipotesi matematiche implicite nella loro derivazione.

8.1 UN PROCESSO CASUALE A TEMPO DISCRETO

Il concetto fondamentale nella rappresentazione matematica dei segnali a energia infinita è quello di *processo casuale* (o *aleatorio*). Nella discussione che stiamo per fare sui processi casuali come modelli dei segnali a tempo discreto e a energia infinita, assumeremo che il lettore abbia familiarità con i principali concetti della teoria della probabilità, come

quelli di variabile casuale, distribuzioni di probabilità, e medie. Per quei lettori che ritenessero tuttavia necessaria una più approfondita conoscenza della teoria della probabilità, rinviando a uno dei testi base elencati nella bibliografia [1-7].

8.1.1 Un esempio semplice: il processo di Bernoulli

Introdurremo il concetto di processo casuale attraverso la discussione di un esempio molto semplice. Supponiamo che una sequenza di numeri sia generata nel modo seguente: ad un dato tempo n viene lanciata una moneta, e, se il risultato è « testa », il valore della sequenza in n è $x(n) = +1$; se il risultato è « croce », il valore è $x(n) = -1$. Una sequenza che potrebbe essere stata generata secondo questo schema è mostrata nella fig. 8.1. Se ammettiamo che questo processo è stato attivo per

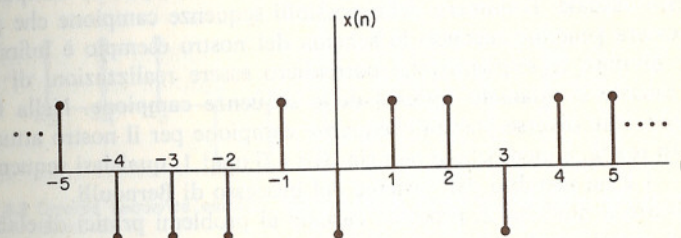


Fig. 8.1 Sequenza di +1 e -1

tutti gli istanti di tempo, cioè per $-\infty < n < \infty$, allora abbiamo una sequenza di durata infinita. Se tentassimo di rappresentare questa sequenza con i metodi dei capitoli precedenti, incontreremmo due difficoltà di base. Innanzitutto è chiaro che la sequenza ha energia infinita, e pertanto non esistono né la trasformata z né quella di Fourier. In secondo luogo la nostra personale esperienza con il lancio delle monete dovrebbe suggerirci che l'unico modo per caratterizzare con precisione la sequenza è quello di tabularne i valori, e, poiché per ipotesi la durata è infinita, anche questo è impossibile. Anche se fosse disponibile un gran numero di campioni passati, sarebbe impossibile determinare con certezza il successivo campione della sequenza. Questa incertezza intrinseca ci spinge a caratterizzare la sequenza in termini di probabilità e a considerare quindi delle medie.

In questo caso supponiamo che, per ogni n , la probabilità di avere testa sia p . Ne segue, per gli assiomi fondamentali della teoria della probabilità, che la probabilità di avere croce deve essere $1 - p$. L' n -mo valore della sequenza, $x(n)$ può allora essere interpretato come un particolare valore di una *variabile casuale* x_n , cioè una funzione dell'uscita o esito del-

l'esperimento del lancio della moneta¹. Specificamente, ogni valore della sequenza può essere visto come il risultato dell'assegnazione di un numero all'uscita dell'esperimento del lancio della moneta. All'evento « è uscito testa » si assegna il valore 1, e analogamente all'evento « è uscito croce » si assegna il valore -1. Poiché questi due eventi mutuamente esclusivi esauriscono l'insieme delle possibili uscite dell'esperimento del lancio della moneta, la variabile casuale x_n può assumere solo uno dei due valori $x(n) = +1$ e $x(n) = -1$. Ad ogni evento si associa un numero che specifica la probabilità che si verifichi quell'evento. Nel nostro esempio la probabilità di avere testa è p e quindi la probabilità che sia $x_n = +1$ è p . Analogamente, poiché la probabilità di avere croce è $1 - p$, tale è anche la probabilità che sia $x_n = -1$.

L'insieme delle variabili casuali $\{x_n\}$ per $-\infty < n < \infty$, unitamente alla descrizione probabilistica delle variabili stesse, definisce un processo casuale². Una data sequenza di valori $\{x(n)\}$, per $-\infty < n < \infty$, è una *realizzazione* del processo casuale e si chiama *sequenza campione* del processo casuale. Il numero delle possibili sequenze campione che potrebbero essere generate secondo lo schema del nostro esempio è infinito. L'insieme di tutte le sequenze che potrebbero essere realizzazioni di un processo casuale è chiamato *insieme* delle sequenze campione. Nella fig. 8.2 sono mostrate diverse possibili sequenze campione per il nostro attuale esempio. In realtà, a meno che p non sia pari a 0 o ad 1, qualsiasi sequenza di $+1$ e -1 è un membro dell'insieme del processo di Bernoulli.

Applicare il modello di processo casuale ai problemi pratici di elaborazione dei segnali implica che una particolare sequenza di dati venga interpretata come un membro dell'insieme delle sequenze campione corrispondenti a un processo casuale. È questo il senso nel quale un processo casuale serve come rappresentazione di un segnale a energia infinita. Dato un segnale a tempo discreto che si assume appartenere a un insieme di sequenze campione, la struttura del corrispondente processo casuale, cioè la legge di probabilità che gli è associata, non è generalmente nota e deve essere dedotta in qualche modo. Può darsi che sia possibile fare delle ipotesi ragionevoli sulla struttura del processo, oppure se ne possono stimare alcune proprietà a partire da un segmento finito di una tipica sequenza campione. Per esempio, sembra plausibile che, nel caso dell'esempio precedente, si possa dedurre la corrispondente legge di probabilità attraverso l'osservazione di un segmento sufficientemente lungo di una delle sequenze di fig. 8.2. Le condizioni sotto le quali ciò può essere fatto saranno discusse nel par. 8.2.2. In ogni caso, per proseguire ulteriormente la discussione di questi modelli di segnali a tempo discreto, è necessario passare a una descrizione più formale di un processo casuale.

¹ Le variabili casuali con questa legge di probabilità sono note come *variabili casuali di Bernoulli*.

² Poiché le variabili casuali sono chiamate variabili casuali di Bernoulli, è ragionevole in questo esempio chiamare *processo casuale di Bernoulli* il corrispondente processo casuale.

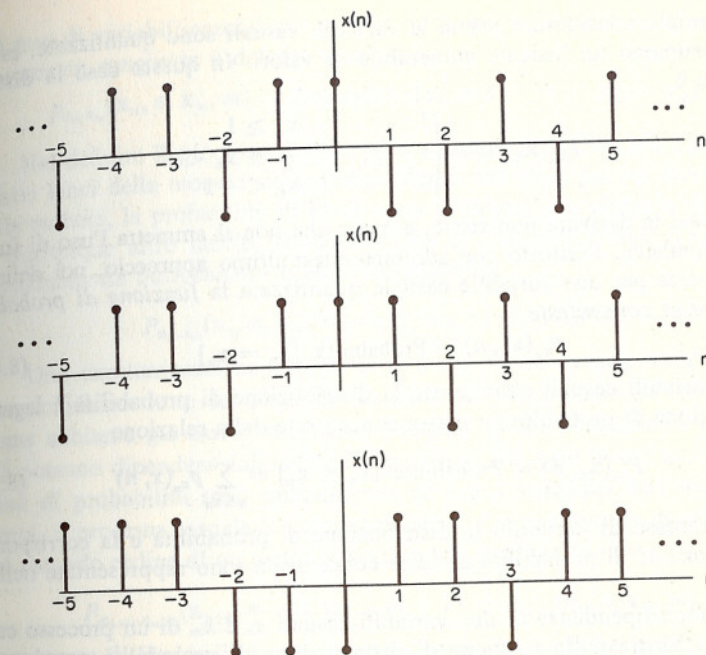


Fig. 8.2 Diverse sequenze estratte dall'insieme di sequenze corrispondenti al processo di Bernoulli

8.1.2 Descrizioni dei processi casuali

Formalmente un processo casuale è un insieme ordinato di variabili casuali $\{x_n\}$ [5-7]. Tale famiglia di variabili casuali è caratterizzata da un insieme di funzioni di distribuzione di probabilità che in generale può dipendere dall'indice n . Quando si usa il concetto di processo come modello per segnali casuali, l'indice n è associato al tempo o anche a qualche altra dimensione fisica. Una singola variabile casuale x_n è descritta dalla funzione di distribuzione di probabilità

$$P_{x_n}(x_n, n) = \text{Probabilità } [x_n \leq x_n] \quad (8.1)$$

dove x_n indica la variabile casuale e x_n un particolare valore di x_n . Se x_n assume un insieme continuo di valori, essa può essere equivalentemente descritta mediante la *funzione di densità di probabilità* definita dalla

$$p_{x_n}(x_n, n) = \frac{\partial P_{x_n}(x_n, n)}{\partial x_n} \quad (8.2)$$

oppure dalla

$$P_{x_n}(x_n, n) = \int_{-\infty}^{x_n} p_{x_n}(x, n) dx \quad (8.3)$$

³ In questo capitolo i caratteri in grassetto sono usati per indicare le variabili correnti delle funzioni di probabilità e *non* vettori o matrici.

Nell'esempio considerato prima le variabili casuali sono quantizzate, esse, cioè, assumono un insieme numerabile di valori. In questo caso la distribuzione è

$$P_{x_n}(x_n, n) = \begin{cases} 1, & x_n \geq 1 \\ 1-p, & -1 \leq x_n < 1 \\ 0, & x_n < -1 \end{cases}$$

In tali casi la derivata non esiste, a meno che non si ammetta l'uso di funzioni impulsive. Piuttosto che adottare quest'ultimo approccio, noi definiremo invece per una variabile casuale quantizzata la *funzione di probabilità a masse concentrate*

$$P_{x_n}(x_n, n) = \text{Probabilità } [x_n = x_n] \quad (8.4)$$

Per le variabili casuali quantizzate la distribuzione di probabilità è legata alla funzione di probabilità a masse concentrate dalla relazione

$$P_{x_n}(x_n, n) = \text{Probabilità } [x_n \leq x_n] = \sum_{x \leq x_n} p_{x_n}(x, n) \quad (8.5)$$

Per il processo di Bernoulli la distribuzione di probabilità e la corrispondente funzione di probabilità a masse concentrate sono rappresentate nella fig. 8.3.

La interdipendenza di due variabili casuali x_n e x_m di un processo casuale è descritta dalla funzione di distribuzione di probabilità congiunta

$$P_{x_n, x_m}(x_n, n, x_m, m) = \text{Probabilità } [x_n \leq x_n \text{ e } x_m \leq x_m] \quad (8.6)$$

o, nel caso di variabili casuali continue, dalla densità di probabilità congiunta

$$p_{x_n, x_m}(x_n, n, x_m, m) = \frac{\partial^2 P_{x_n, x_m}(x_n, n, x_m, m)}{\partial x_n \partial x_m} \quad (8.7)$$

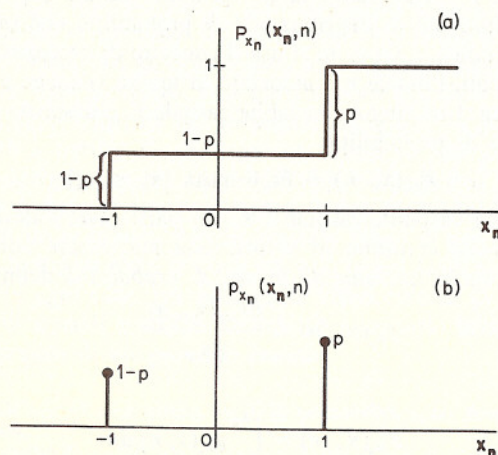


Fig. 8.3 (a) Distribuzione di probabilità di una variabile casuale di Bernoulli; (b) corrispondente funzione di probabilità a masse concentrate

Nel caso di variabili casuali quantizzate, la funzione di probabilità a masse concentrate congiunta è definita come

$$P_{x_n, x_m}(x_n, n, x_m, m) = \text{Probabilità } [x_n = x_n \text{ e } x_m = x_m] \quad (8.8)$$

Nel definire il processo di Bernoulli, abbiamo fatto l'ipotesi che i successivi lanci della moneta siano indipendenti, che cioè, per un dato lancio della moneta, la probabilità di avere testa (o croce) non dipenda dall'esito di qualunque altro lancio. In questo caso le variabili casuali $\{x_n\}$ sono *statisticamente indipendenti* e cioè, formalmente,

$$P_{x_n, x_m}(x_n, n, x_m, m) = P_{x_n}(x_n, n) \cdot P_{x_m}(x_m, m)$$

Una caratterizzazione completa di un processo casuale richiede che si specifichino tutte le possibili distribuzioni di probabilità congiunte. Come abbiamo già osservato, queste funzioni di distribuzione di probabilità possono dipendere dall'indice temporale n . Nel caso in cui tutte le funzioni di probabilità sono indipendenti da una traslazione dell'origine dei tempi, il processo casuale è detto *stazionario*. Per esempio, la distribuzione del secondo ordine di un processo stazionario soddisfa la relazione

$$P_{x_{n+k}, x_{m+k}}(x_{n+k}, n+k, x_{m+k}, m+k) = P_{x_n, x_m}(x_n, n, x_m, m) \quad (8.9)$$

Il processo di Bernoulli è un esempio di processo stazionario in quanto si è fatta l'ipotesi che nel lanciare la moneta la probabilità di avere testa è sempre uguale a p e che ogni variabile aleatoria è indipendente da tutte le altre.

In molte applicazioni proprie dell'elaborazione numerica dei segnali, i processi casuali servono come modelli dei segnali, nel senso che un particolare segnale a energia infinita può essere considerato una sequenza campione di un processo casuale. Benché i dettagli di tali segnali siano imprevedibili (rendendo inadeguato il nostro precedente approccio alla rappresentazione dei segnali), certe proprietà medie dell'insieme possono tuttavia essere previste se è nota la legge di probabilità del processo. Queste proprietà medie servono spesso a dare una caratterizzazione utile, anche se incompleta, di quei segnali per i quali non è applicabile l'approccio deterministico.

8.2 MEDIE

Si sa che è spesso utile caratterizzare una variabile casuale mediante dei valori medi quali la « media » (propriamente detta) e la varianza. Poiché un processo casuale è un insieme ordinato di variabili casuali, si può analogamente caratterizzare il processo mediante medie statistiche delle variabili che lo costituiscono. Tali medie si chiamano *medie d'insieme*.

8.2.1 Definizioni

Il valore medio o la media di un processo è definito come

$$m_{x_n} = E[x_n] = \int_{-\infty}^{\infty} x p_{x_n}(x, n) dx \quad (8.10)$$

dove l'operatore E indica appunto media o « valore atteso ». Notiamo che in generale la media (il valore atteso) può dipendere da n .

In generale, se $g(\cdot)$ è una funzione a un sol valore, allora anche $g(x_n)$ è una variabile casuale, e l'insieme delle variabili casuali $g(x_n)$ definisce un nuovo processo casuale. Per calcolare le medie di questo nuovo processo si possono derivare le distribuzioni di probabilità delle nuove variabili casuali. Alternativamente, si può dimostrare che

$$E[g(x_n)] = \int_{-\infty}^{\infty} g(x) p_{x_n}(x, n) dx \quad (8.11)$$

Se le variabili aleatorie sono quantizzate, gli integrali diventano sommatorie su tutti i possibili valori delle variabili:

$$E[g(x_n)] = \sum_x g(x) p_{x_n}(x, n) \quad (8.12)$$

Nei casi in cui si è interessati alla interdipendenza fra due (o più) segnali a energia infinita, cioè fra due processi casuali, è necessario operare su due insiemi di variabili casuali $\{x_n\}$ ed $\{y_m\}$. Per esempio, il valore atteso di una funzione di due variabili casuali è per definizione

$$E[g(x_n, y_m)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) p_{x_n, y_m}(x, n, y, m) dx dy \quad (8.13)$$

dove $p_{x_n, y_m}(x_n, n, y_m, m)$ è la densità di probabilità congiunta delle variabili casuali x_n e y_m .

Esistono diverse semplici proprietà delle medie che torneranno utili nella discussione che stiamo per fare. In particolare, si dimostra facilmente che:

1. $E[x_n + y_m] = E[x_n] + E[y_m]$, cioè la media di una somma è uguale alla somma delle medie.
2. $E[a \cdot x_n] = a \cdot E[x_n]$, cioè la media di una costante per x_n è uguale alla costante per la media di x_n .

In generale la media del prodotto di due variabili aleatorie non è uguale al prodotto delle medie. Se ciò tuttavia si verifica, allora le due variabili aleatorie si dicono *linearmente indipendenti* o *scorrelate*. In tal caso quindi

$$E[x_n y_m] = E[x_n] \cdot E[y_m] \quad (8.14)$$

Si verifica facilmente dalla (8.13) che una condizione sufficiente per l'indipendenza lineare è

$$p_{x_n y_m}(x_n, n, y_m, m) = p_{x_n}(x_n, n) \cdot p_{y_m}(y_m, m) \quad (8.15)$$

Questa condizione di indipendenza è più forte tuttavia della (8.14). Come già affermato in precedenza, le variabili casuali che soddisfano la (8.15) si dicono *statisticamente indipendenti*. Se la (8.15) vale per tutti i valori di n ed m , allora i processi casuali $\{x_n\}$ e $\{y_m\}$ si dicono statisticamente indipendenti. I processi casuali statisticamente indipendenti sono anche linearmente indipendenti, mentre non vale il viceversa, in quanto l'indipendenza lineare non implica quella statistica.

Dalle espressioni (8.11) - (8.13) si vede che le medie sono generalmente funzioni del tempo. Tuttavia, nel caso dei processi stazionari ciò non è vero. Pertanto il valor medio è lo stesso per tutte le variabili casuali che costituiscono il processo e quindi la media di un processo stazionario è una costante che indicheremo semplicemente con m_x .

Per un processo casuale, oltre al valor medio, definito dalla (8.10), esistono diverse altre medie che sono particolarmente importanti nel contesto della elaborazione numerica dei segnali. Queste sono definite qui di seguito.

Per comodità di notazione assumeremo che le distribuzioni di probabilità siano continue. Le corrispondenti definizioni per i processi casuali quantizzati si ottengono applicando la (8.12).

Il *valore quadratico medio* di x_n è la media di x_n^2 , cioè

$$E[x_n^2] = \text{media quadratica} = \int_{-\infty}^{\infty} x^2 p_{x_n}(x, n) dx \quad (8.16)$$

Il valore quadratico medio viene talora chiamato anche *potenza media*.

La *varianza* di x_n è il valore quadratico medio di $[x_n - m_{x_n}]$, cioè

$$\text{varianza} = E[(x_n - m_{x_n})^2] = \sigma_{x_n}^2 \quad (8.17)$$

Poiché la media di una somma è la somma delle medie, si dimostra facilmente che la (8.17) si può scrivere

$$\begin{aligned} \text{varianza} &= E[x_n^2] - m_{x_n}^2 \\ &= \text{media quadratica} - (\text{media})^2 \end{aligned} \quad (8.18)$$

In generale il valore quadratico medio e la varianza sono funzioni del tempo; tuttavia per i processi stazionari essi sono delle costanti.

La media, la media quadratica e la varianza sono medie semplici che forniscono solo una piccola quantità di informazione sul processo. Una media più utile è la *sequenza di autocorrelazione* che è definita come

$$\begin{aligned} \phi_{xx}(n, m) &= E[x_n x_m^*] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_n x_m^* p_{x_n, x_m}(x_n, n, x_m, m) dx_n dx_m \end{aligned} \quad (8.19)$$

dove l'asterisco indica complesso coniugato. La sequenza di autocovarianza di un processo casuale è definita come

$$\gamma_{xx}(n, m) = E[(x_n - m_{x_n})(x_m - m_{x_m})^*] \quad (8.20)$$

che può essere anche scritta come

$$\gamma_{xx}(n, m) = \phi_{xx}(n, m) - m_{x_n} m_{x_m} \quad (8.21)$$

Si noti che in generale sia l'autocorrelazione che l'autocovarianza sono sequenze bidimensionali.

L'autocorrelazione è una misura della dipendenza tra i valori di un processo casuale in tempi diversi. In questo senso essa descrive la variazione nel tempo di un segnale casuale. Una misura della dipendenza tra due diversi segnali casuali si ottiene con la sequenza di correlazione incrociata. Se $\{x_n\}$ e $\{y_m\}$ sono due processi casuali, la loro correlazione incrociata è

$$\begin{aligned}\phi_{xy}(n, m) &= E[x_n y_m^*] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy^* p_{x_n, y_m}(x, n, y, m) dx dy\end{aligned}\quad (8.22)$$

dove $p_{x_n, y_m}(x, n, y, m)$ è la densità di probabilità congiunta di x_n e y_m . La funzione di covarianza incrociata è definita come

$$\begin{aligned}\gamma_{xy}(n, m) &= E[(x_n - m_{x_n})(y_m - m_{y_m})^*] \\ &= \phi_{xy}(n, m) - m_{x_n} m_{y_m}^*\end{aligned}\quad (8.23)$$

Come si è già detto, le proprietà statistiche di un processo casuale variano generalmente col tempo. Tuttavia, un processo stazionario è caratterizzato da una condizione di equilibrio per cui le proprietà statistiche sono invarianti a una traslazione dell'origine dei tempi. Ciò implica che la distribuzione di probabilità del primo ordine è indipendente dal tempo. Analogamente anche tutte le funzioni di probabilità congiunta sono invarianti a una traslazione dell'origine dei tempi; in particolare la distribuzione di probabilità congiunta del secondo ordine soddisfa la (8.9). Dalla (8.9) segue che la distribuzione congiunta del secondo ordine dipende solo dalla differenza nei tempi $m-n$. Le medie del primo ordine, come la media e la varianza, sono quindi indipendenti dal tempo; le medie del secondo ordine, come l'autocorrelazione $\phi_{xx}(n, m)$, sono dipendenti dalla differenza temporale $m-n$. Pertanto, per un processo stazionario possiamo scrivere

$$m_x = E[x_n] \quad (8.24)$$

$$\sigma_x^2 = E[(x_n - m_x)^2] \quad (8.25)$$

indipendenti da n , e, se indichiamo ora con m la differenza temporale,

$$\phi_{xx}(n, n+m) = \phi_{xx}(m) = E[x_n x_{n+m}^*] \quad (8.26)$$

Vale a dire che la sequenza di autocorrelazione di un processo casuale stazionario è una sequenza monodimensionale, in quanto è una funzione della differenza temporale m .

In molte situazioni si incontrano processi casuali che non sono stazionari in senso stretto; vale a dire che le loro distribuzioni di probabilità non sono invarianti nel tempo, eppure la media è costante e la sequenza di autocorrelazione soddisfa la (8.26). Tali processi casuali si dicono *stazionari in senso lato* [5].

ESEMPIO. Come esempio di descrizione di un processo casuale mediante medie, consideriamo ancora il semplice processo di Bernoulli. Per cominciare, osserviamo che il processo è stazionario in quanto si è assunto che le probabilità di $+1$ e -1 sono indipendenti dal tempo e che le variabili casuali $\{x_n\}$ sono statisticamente indipendenti. Usando la (8.12) troviamo che la media è

$$\begin{aligned}m_x &= (+1) \cdot \text{Probabilità } [x_n = +1] + (-1) \cdot \text{Probabilità } [x_n = -1] \\ &= +1 \cdot p + (-1) \cdot (1-p) \\ &= (2p-1)\end{aligned}$$

e il valore quadratico medio è

$$\begin{aligned}\text{media quadratica} &= (+1)^2 \cdot \text{Probabilità } [x_n = +1] \\ &\quad + (-1)^2 \cdot \text{Probabilità } [x_n = -1] \\ &= (+1)^2 p + (-1)^2 (1-p) \\ &= 1\end{aligned}$$

Pertanto la varianza è

$$\sigma_x^2 = 1 - (2p-1)^2 = 4p(1-p)$$

Avendo assunto l'indipendenza statistica, la sequenza di autocorrelazione è

$$\phi_{xx}(m) = \begin{cases} E[x_n^2] = 1, & m = 0 \\ E[x_n] \cdot E[x_{n+m}] = m_x^2, & m \neq 0 \end{cases}$$

In particolare, se $p = 1/2$, allora $m_x = 0$ e

$$\phi_{xx}(m) = \delta(m)$$

In generale, una tale sequenza di autocorrelazione si ottiene tutte le volte che le variabili casuali di un processo casuale sono linearmente indipendenti. Simili processi (chiamati *rumore bianco*) giocano un ruolo importante in molti problemi di elaborazione dei segnali.

8.2.2 Medie temporali

Come già accennato varie volte, la nozione di un insieme di segnali a energia infinita è un concetto matematico utile, nel contesto della elaborazione dei segnali, in quanto ci consente di usare la teoria della probabilità per rappresentare segnali a energia infinita. Tuttavia, in pratica, si preferirebbe avere a che fare con una singola sequenza piuttosto che con un insieme infinito di sequenze. Per esempio, sarebbe desiderabile poter ricavare la legge di probabilità o certe medie relative alla rappresentazione di un processo casuale, attraverso misure fatte sopra un singolo elemento dell'insieme. Per il processo di Bernoulli, per esempio, ricordiamo che le distribuzioni di probabilità sono indipendenti dal tempo, e pertanto si è portati intuitivamente ad ammettere che le percentuali di $+1$ e -1 in un tratto abbastanza lungo di una singola sequenza campione debbano essere prossime rispettivamente a p e $1-p$. Analogamente la media aritmetica di un gran numero di campioni di una singola sequenza dovrebbe approssimare il valor medio del processo. Allo scopo di formalizzare que-

ste nozioni intuitive, definiamo la media temporale di un processo casuale come

$$\langle x_n \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x_n \quad (8.27)$$

Analogamente, la sequenza di autocorrelazione nel tempo è definita come

$$\langle x_n x_{n+m} \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x_n x_{n+m}^* \quad (8.28)$$

Si può dimostrare che i limiti ora definiti esistono se $\{x_n\}$ è un processo stazionario con media finita. La dimostrazione di questo risultato andrebbe tuttavia ben oltre i limiti di questa nostra discussione. Come risulta dalle (8.27) e (8.28), queste medie nel tempo sono funzioni di un insieme infinito di variabili casuali e quindi vanno considerate esse stesse come variabili casuali. Tuttavia, sotto una condizione nota come condizione di *ergodicità*, le medie temporali (8.27) e (8.28) sono delle costanti, nel senso che le medie temporali di quasi tutte le possibili sequenze campione sono uguali alla stessa costante. Esse sono inoltre uguali alle corrispondenti medie d'insieme⁴. Vale a dire che, per ogni singola sequenza campione $\{x(n)\}$, per $-\infty < n < +\infty$,

$$\langle x(n) \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n) = E[x_n] = m_x \quad (8.29)$$

e

$$\begin{aligned} \langle x(n)x^*(n+m) \rangle \\ = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n)x^*(n+m) = E[x_n x_{n+m}^*] = \phi_{xx}(m) \end{aligned} \quad (8.30)$$

L'operatore di media temporale $\langle \rangle$ ha le stesse proprietà dell'operatore di media d'insieme $E[\]$. Pertanto noi spesso non ci preoccupiamo molto di distinguere tra la variabile casuale x_n e il suo valore in una sequenza campione, $x(n)$. Per esempio, l'espressione $E[x(n)]$ dovrà essere interpretata come $E[x_n] = \langle x(n) \rangle$. In generale un processo casuale per il quale le medie temporali sono uguali alle medie d'insieme è chiamato *processo ergodico* [5, 6].

In pratica si ammette di solito che una data sequenza è una sequenza campione di un processo casuale ergodico. Pertanto le medie sono calcolabili a partire da una singola sequenza a energia infinita. Ovviamente non è possibile in generale calcolare proprio i limiti delle (8.29) e (8.30), ma si calcolano piuttosto le quantità

$$\langle x(n) \rangle_N = \frac{1}{2N+1} \sum_{n=-N}^N x(n) \quad (8.31)$$

⁴ Un'affermazione più precisa è che le variabili casuali $\langle x_n \rangle$ e $\langle x_n x_{n+m}^* \rangle$ hanno medie uguali rispettivamente a m_x e $\phi_{xx}(m)$, e le loro varianze sono zero [6].

$$\langle x(n)x(n+m) \rangle_N = \frac{1}{2N+1} \sum_{n=-N}^N x(n)x^*(n+m) \quad (8.32)$$

o altre quantità simili che sono delle *stime* della media e dell'autocorrelazione [8, 9]. La stima delle medie di un processo casuale a partire da un segmento finito di dati è un problema di statistica, che noi prenderemo in esame nel cap. 11.

8.3 RAPPRESENTAZIONI IN FREQUENZA DEI SEGNALE A ENERGIA INFINITA

Sebbene per un segnale a energia infinita la trasformata z non esista, le sue sequenze di autocorrelazione e autocovarianza sono sequenze aperiodiche per le quali invece la trasformata z e la trasformata di Fourier spesso esistono. Vedremo nel prossimo paragrafo che la rappresentazione spettrale (cioè in frequenza) di queste medie gioca un ruolo importante nella descrizione delle relazioni ingresso-uscita di un sistema lineare tempo-invariante quando l'ingresso è un segnale a energia infinita. È pertanto utile considerare le proprietà delle sequenze di correlazione e covarianza e le loro corrispondenti trasformate z .

8.3.1 Proprietà delle sequenze di correlazione e covarianza

Esistono numerose e utili proprietà delle funzioni di correlazione e covarianza che derivano in modo semplice dalle loro definizioni. Queste proprietà sono elencate qui di seguito onde poter essere utilizzate facilmente in futuro. La dimostrazione della validità di alcune di queste proprietà è considerata nel probl. 6 di questo stesso capitolo.

Si considerino due processi casuali stazionari e reali, $\{x_n\}$ e $\{y_n\}$, con autocorrelazione, autocovarianza, correlazione incrociata e covarianza incrociata date rispettivamente da

$$\phi_{xx}(m) = E[x_n x_{n+m}] \quad (8.33)$$

$$\gamma_{xx}(m) = E[(x_n - m_x)(x_{n+m} - m_x)] \quad (8.34)$$

$$\phi_{xy}(m) = E[x_n y_{n+m}] \quad (8.35)$$

$$\gamma_{xy}(m) = E[(x_n - m_x)(y_{n+m} - m_y)] \quad (8.36)$$

dove m_x ed m_y sono le medie dei due processi. Dalle definizioni, con semplici passaggi, si ricavano facilmente le seguenti proprietà.

Proprietà 1:

$$\gamma_{xx}(m) = \phi_{xx}(m) - m_x^2 \quad (8.37a)$$

$$\gamma_{xy}(m) = \phi_{xy}(m) - m_x m_y \quad (8.37b)$$

Questi risultati sono una diretta conseguenza delle (8.21) e (8.23), e indicano che le sequenze di correlazione e covarianza sono identiche se $m_x = 0$.

Proprietà 2:

$$\phi_{xx}(0) = E[x_n^2] = \text{valore quadratico medio} \quad (8.38a)$$

$$\gamma_{xx}(0) = \sigma_x^2 = \text{varianza} \quad (8.38b)$$

Proprietà 3:

$$\phi_{xx}(m) = \phi_{xx}(-m) \quad (8.39a)$$

$$\gamma_{xx}(m) = \gamma_{xx}(-m) \quad (8.39b)$$

$$\phi_{xy}(m) = \phi_{yx}(-m) \quad (8.39c)$$

$$\gamma_{xy}(m) = \gamma_{yx}(-m) \quad (8.39d)$$

Proprietà 4:

$$|\phi_{xy}(m)| \leq [\phi_{xx}(0)\phi_{yy}(0)]^{1/2} \quad (8.40a)$$

$$|\gamma_{xy}(m)| \leq [\gamma_{xx}(0)\gamma_{yy}(0)]^{1/2} \quad (8.40b)$$

In particolare, quindi,

$$|\phi_{xx}(m)| \leq \phi_{xx}(0) \quad (8.41a)$$

$$|\gamma_{xx}(m)| \leq \gamma_{xx}(0) \quad (8.41b)$$

Proprietà 5: Se $y_n = x_{n-n_0}$, allora

$$\phi_{yy}(m) = \phi_{xx}(m) \quad (8.42a)$$

$$\gamma_{yy}(m) = \gamma_{xx}(m) \quad (8.42b)$$

Proprietà 6: Per molti processi casuali le variabili casuali che li costituiscono diventano meno correlate quanto più sono separate nel tempo. Ciò corrisponde alle seguenti relazioni

$$\lim_{m \rightarrow \infty} \phi_{xx}(im) = (E[x_n])^2 = m_x^2 \quad (8.43a)$$

$$\lim_{m \rightarrow \infty} \gamma_{xx}(im) = 0 \quad (8.43b)$$

$$\lim_{m \rightarrow \infty} \phi_{xy}(im) = m_x m_y \quad (8.43c)$$

$$\lim_{m \rightarrow \infty} \gamma_{xy}(im) = 0 \quad (8.43d)$$

L'essenza di questi risultati è che: la correlazione e la covarianza sono sequenze aperiodiche che tendono a smorzarsi all'aumentare di m . Pertanto è spesso possibile rappresentare queste sequenze per mezzo delle loro trasformate z .

8.3.2 Rappresentazioni con trasformate z

Indichiamo con $\Phi_{xx}(z)$, $\Gamma_{xx}(z)$, $\Phi_{xy}(z)$ e $\Gamma_{xy}(z)$ le trasformate z rispettivamente di $\phi_{xx}(m)$, $\gamma_{xx}(m)$, $\phi_{xy}(m)$ e $\gamma_{xy}(m)$. Per le (8.43a) e (8.43c), osserviamo subito che le trasformate z di $\phi_{xx}(m)$ e $\phi_{xy}(m)$ esistono solo se $m_x = 0$, nel qual caso $\Phi_{xx}(z) = \Gamma_{xx}(z)$ e $\Phi_{xy}(z) = \Gamma_{xy}(z)$. Numerose altre proprietà delle trasformate z si ricavano dalle proprietà delle sequenze di correlazione e covarianza riassunte nel par. 8.3.1. Queste proprietà delle trasfor-

mate z sono riassunte qui di seguito, mentre le loro dimostrazioni sono prese in esame nel probl. 6 di questo capitolo.

Proprietà 1:

$$\sigma_x^2 = \frac{1}{2\pi j} \oint_C \Gamma_{xx}(z) z^{-1} dz \quad (8.44)$$

dove C è un percorso chiuso nella regione di convergenza di $\Gamma_{xx}(z)$.

Proprietà 2:

$$\Gamma_{xx}(z) = \Gamma_{xx}(1/z) \quad (8.45a)$$

$$\Gamma_{xy}(z) = \Gamma_{yx}^*(1/z^*) \quad (8.45b)$$

Le (8.45) seguono direttamente dalla proprietà 3 del par. 8.3.1. Ne consegue che la regione di convergenza di $\Gamma_{xx}(z)$ deve essere della forma

$$R_a < |z| < \frac{1}{R_a}$$

Inoltre, poiché $\gamma_{xx}(m)$ diventa zero per $m = \infty$, la regione di convergenza deve includere il circolo unitario, dove cioè essere $0 < R_a < 1$. Nel caso importante in cui $\Gamma_{xx}(z)$ sia una funzione razionale di z , ciò implica che i suoi poli e zeri devono essere a due a due reciproci complessi coniugati come indicato nella fig. 8.4.

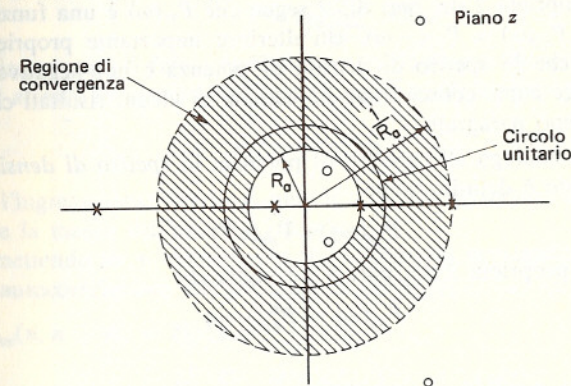


Fig. 8.4 Regione di convergenza e posizione di poli e zeri di una tipica trasformata z di una sequenza di covarianza

8.3.3 Spettro di potenza

Poiché la regione di convergenza contiene il circolo unitario, la (8.44) si può scrivere

$$\sigma_x^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\omega) d\omega \quad (8.46)$$

dove si è posto

$$P_{xx}(\omega) = \Gamma_{xx}(e^{j\omega}) \quad (8.47)$$

Ricordiamo che, quando $m_x = 0$, la varianza è uguale alla media quadratica o potenza media del segnale. Pertanto l'area sottesa da $P_{xx}(\omega)$ per $-\pi \leq \omega \leq \pi$ è proporzionale alla potenza media del segnale. In effetti, come vedremo nel prossimo paragrafo, l'integrale di $P_{xx}(\omega)$ sopra una certa banda di frequenze è proporzionale alla potenza del segnale in quella banda. Per queste ragioni la funzione $P_{xx}(\omega)$ viene chiamata lo *spettro di densità di potenza*, o semplicemente lo *spettro* [9]. Osserviamo che è anche frequente definire lo spettro di potenza come la trasformata di Fourier della sequenza di autocorrelazione piuttosto che della autocovarianza [5]. Ciò comporta qualche difficoltà quando $m_x \neq 0$, poiché $\phi_{xx}(m) \rightarrow m_x^2$ quando $m \rightarrow \infty$. Ne segue infatti che la trasformata di Fourier della sequenza di autocorrelazione non esiste se $m_x \neq 0$, a meno che non si voglia estendere la nostra definizione di trasformata di Fourier in modo da consentire la presenza di un impulso nello spettro di potenza in $\omega = 0$. Ma dal momento che in questo libro si è evitato l'uso degli impulsi, preferiamo definire lo spettro di potenza come nella (8.48). Osserviamo che quando $m_x = 0$, le sequenze di autocorrelazione e autocovarianza sono identiche, e tali pertanto sono anche le loro trasformate di Fourier.

Dalla proprietà 2 del par. 8.3.2 segue che $P_{xx}(\omega)$ è una funzione simmetrica, cioè $P_{xx}(\omega) = P_{xx}(-\omega)$. Un'ulteriore importante proprietà consiste nel fatto che lo spettro di densità di potenza è non-negativo. Questa proprietà nasce come conseguenza immediata di alcuni risultati che vedremo nel prossimo paragrafo.

In modo analogo allo spettro di potenza, lo *spettro di densità di potenza incrociato* è definito come

$$P_{xy}(\omega) = \Gamma_{xy}(e^{j\omega}) \quad (8.48)$$

Ancora dalla proprietà 2 del par. 8.3.2 segue che

$$P_{xy}(\omega) = P_{yx}^*(-\omega) \quad (8.49)$$

8.4 RISPOSTA DEI SISTEMI LINEARI A SEGNALI CASUALI

Nei capitoli precedenti è stata sviluppata la teoria dei sistemi lineari a tempo discreto per il caso in cui l'ingresso è una funzione nota del tempo. Le discussioni svolte hanno mostrato come la nozione di risposta in frequenza di un sistema lineare invariante alla traslazione e la rappresentazione nel dominio della frequenza di un segnale a tempo discreto sono concetti essenziali nell'elaborazione numerica dei segnali. In questo paragrafo svilupperemo i concetti corrispondenti per il caso dei segnali a energia infinita, cioè per i segnali rappresentanti mediante processi casuali.

Si consideri un sistema lineare stabile invariante alla traslazione con risposta all'impulso $h(n)$. Sia $x(n)$ una sequenza di ingresso reale che è anche una sequenza campione di un processo casuale a tempo discreto stazionario in senso lato. Allora l'uscita del sistema lineare è una sequenza

campione di un processo casuale di uscita che è legato al processo di ingresso dalla trasformazione lineare

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)x(k) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) \quad (8.50)$$

Come si è mostrato a suo tempo, poiché il sistema è stabile, $y(n)$ sarà limitata se $x(n)$ è limitata. Dimosteremo più avanti che se l'ingresso è stazionario, tale è anche l'uscita. Il segnale di ingresso può essere caratterizzato dalla sua media m_x e la sua funzione di autocorrelazione $\phi_{xx}(m)$, e da altre informazioni circa le sue distribuzioni di probabilità del primo e del secondo ordine. Per caratterizzare il processo casuale di uscita $\{y_n\}$ si desiderano avere informazioni analoghe. Per molte applicazioni è tuttavia sufficiente caratterizzare sia l'ingresso che l'uscita in termini di medie semplici, quali il valor medio, la varianza e l'autocorrelazione. Pertanto deriveremo ora le relazioni ingresso-uscita fra queste quantità.

La media del processo di uscita è

$$\begin{aligned} m_y &= E[y(n)] = \sum_{k=-\infty}^{\infty} h(k)E[x(n-k)] \\ &= m_x \sum_{k=-\infty}^{\infty} h(k) \end{aligned} \quad (8.50)$$

dove abbiamo usato il fatto che il valore atteso di una somma è la somma dei valori attesi⁵. In termini della funzione di trasferimento possiamo scrivere

$$m_y = H(e^{j0})m_x \quad (8.51)$$

Essendo l'ingresso stazionario e quindi a media costante, ne deduciamo che anche la media dell'uscita è costante.

Ammettendo temporaneamente che l'uscita sia non-stazionaria, la funzione di autocorrelazione del processo di uscita è

$$\begin{aligned} \phi_{yy}(n, n+m) &= E[y(n)y(n+m)] \\ &= E\left[\sum_{k=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(k)h(r)x(n-k)x(n+m-r)\right] \\ &= \sum_{k=-\infty}^{\infty} h(k) \sum_{r=-\infty}^{\infty} h(r)E[x(n-k)x(n+m-r)] \end{aligned}$$

Poiché $x(n)$ è per ipotesi stazionario, $E[x(n-k)x(n+m-r)]$ dipende solo dalla differenza temporale $m+k-r$. Pertanto

$$\phi_{yy}(n, n+m) = \sum_{k=-\infty}^{\infty} h(k) \sum_{r=-\infty}^{\infty} h(r)\phi_{xx}(m+k-r) = \phi_{yy}(m) \quad (8.52)$$

e cioè, anche la sequenza di autocorrelazione dell'uscita dipende solo dalla differenza temporale m . Pertanto, per un sistema lineare invariante alla traslazione eccitato da un ingresso stazionario, anche l'uscita è stazionaria.

⁵ Si noti che abbiamo cominciato ad essere meno scrupolosi nel distinguere fra la variabile casuale e il suo valore.

Con la sostituzione $l = r - k$, la (8.52) si può scrivere

$$\begin{aligned}\phi_{yy}(m) &= \sum_{l=-\infty}^{\infty} \phi_{xx}(m-l) \sum_{k=-\infty}^{\infty} h(k)h(l+k) \\ &= \sum_{k=-\infty}^{\infty} \phi_{xx}(m-l)v(l)\end{aligned}\quad (8.53)$$

dove si è posto

$$v(l) = \sum_{k=-\infty}^{\infty} h(k)h(l+k) \quad (8.54)$$

Una sequenza della forma di $v(l)$ è spesso chiamata *sequenza di autocorrelazione aperiodica* o semplicemente la *sequenza di autocorrelazione* di $h(n)$. Va sottolineato che $v(l)$ è l'autocorrelazione di una sequenza aperiodica, cioè a energia finita, e non va confusa con l'autocorrelazione di una sequenza a energia infinita. In effetti, si può vedere che $v(l)$ non è altro che la convoluzione discreta di $h(n)$ con $h(-n)$. La (8.53), allora, si può interpretare nel senso che l'autocorrelazione dell'uscita di un sistema lineare è la convoluzione dell'autocorrelazione dell'ingresso con l'autocorrelazione della risposta all'impulso del sistema.

La (8.53) suggerisce che la trasformata z può essere utile per caratterizzare la risposta di un sistema lineare invariante nel tempo a un ingresso a energia infinita. Si assuma per convenienza che $m_x = 0$, per cui le sequenze di autocorrelazione e autocovarianza sono identiche. Dalle (8.53) e (8.54) si ha

$$\begin{aligned}\Phi_{yy}(z) &= V(z)\Phi_{xx}(z) \\ &= H(z)H(z^{-1})\Phi_{xx}(z)\end{aligned}\quad (8.55)$$

In termini dello spettro di densità di potenza, la (8.55) diventa

$$P_{yy}(\omega) = |H(e^{j\omega})|^2 P_{xx}(\omega) \quad (8.56)$$

È la (8.56) che motiva l'uso del termine « spettro di densità di potenza ». Per vederlo, ammettiamo che sia $m_x = 0$, così che, per la (8.51), anche $m_y = 0$. Pertanto

$$\begin{aligned}\phi_{yy}(0) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{yy}(\omega) d\omega \\ &= \text{potenza media totale in uscita}\end{aligned}\quad (8.57)$$

Sostituendo la (8.56) nella (8.57) si ha

$$\phi_{yy}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 P_{xx}(\omega) d\omega \quad (8.58)$$

Si supponga ora che $H(e^{j\omega})$ sia un filtro passa-banda ideale, come mostrato nella fig. 8.5. Ricordiamo che $\phi_{xx}(m)$ è una sequenza pari, per cui

$$P_{xx}(\omega) = P_{xx}(-\omega)$$

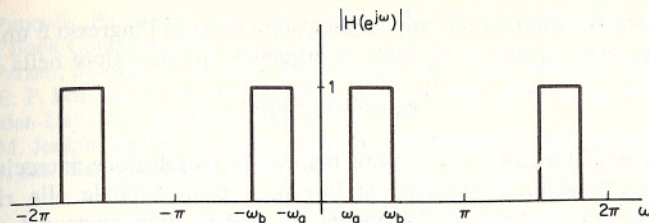


Fig. 8.5 Risposta in frequenza di un filtro passa-banda ideale

Analogamente, anche $|H(e^{j\omega})|^2$ è una funzione pari di ω . Perciò possiamo scrivere

$$\begin{aligned}\phi_{yy}(0) &= \text{potenza media in uscita} \\ &= \frac{1}{\pi} \int_{\omega_a}^{\omega_b} P_{xx}(\omega) d\omega\end{aligned}\quad (8.59)$$

Pertanto l'area sottesa da $P_{xx}(\omega)$ tra ω_a e ω_b rappresenta il valore quadratico medio dell'ingresso in quella banda di frequenza. Osserviamo che la potenza di uscita deve restare non negativa, per cui

$$\lim_{(\omega_b - \omega_a) \rightarrow 0} \phi_{yy}(0) \geq 0$$

Questo risultato, insieme alla (8.59), implica che

$$P_{xx}(\omega) \geq 0 \quad (8.60)$$

Se ne conclude pertanto che lo spettro di densità di potenza di un segnale reale è reale, pari e positivo.

Un altro risultato interessante riguarda la correlazione incrociata tra l'ingresso e l'uscita di un sistema lineare invariante nel tempo:

$$\begin{aligned}\phi_{xy}(m) &= E[x(n)y(n+m)] \\ &= E\left[x(n) \sum_{k=-\infty}^{\infty} h(k)x(n+m-k)\right] \\ &= \sum_{k=-\infty}^{\infty} h(k)\phi_{xx}(m-k)\end{aligned}\quad (8.61)$$

In questo caso osserviamo che la correlazione incrociata fra l'ingresso e l'uscita è la convoluzione fra la risposta all'impulso e la sequenza di autocorrelazione dell'ingresso.

Se si assume $m_x = 0$, in modo che esistano le trasformate z , possiamo scrivere

$$\Phi_{xy}(z) = H(z)\Phi_{xx}(z) \quad (8.62)$$

oppure, in termini degli spettri di potenza,

$$P_{xy}(\omega) = H(e^{j\omega})P_{xx}(\omega) \quad (8.63)$$

Questo risultato ha un'utile applicazione quando l'ingresso è un rumore bianco, cioè $\phi_{xx}(m) = \sigma_x^2 \delta(m)$. Sostituendo questo valore nella (8.61), abbiamo

$$\phi_{xy}(m) = \sigma_x^2 h(m) \quad (8.64)$$

e quindi, se l'ingresso è un rumore bianco, la correlazione incrociata fra l'ingresso e l'uscita di un sistema lineare è proporzionale alla risposta all'impulso del sistema. Analogamente, lo spettro di potenza del rumore bianco d'ingresso è

$$P_{xx}(\omega) = \sigma_x^2, \quad -\pi \leq \omega \leq \pi$$

e pertanto dalla (8.63) si deduce

$$P_{xy}(\omega) = \sigma_x^2 H(e^{j\omega}) \quad (8.65)$$

e quindi lo spettro di potenza incrociato risulta in questo caso proporzionale alla risposta in frequenza del sistema. Le (8.64) e (8.65) possono servire come base per stimare la risposta all'impulso o la risposta in frequenza di un sistema lineare invariante nel tempo quando è possibile osservare l'uscita del sistema in risposta a un rumore bianco di ingresso.

SOMMARIO

In questo capitolo abbiamo cercato di mostrare come si può usare il concetto di processo casuale per rappresentare quei segnali a tempo discreto per i quali gli strumenti matematici tipo trasformate di Fourier non sono direttamente applicabili. Non si è avuto l'intenzione di sostituire con questo capitolo un corso formale in probabilità e processi aleatori. Pertanto i nostri sforzi, più che su discussioni rigorose, si sono concentrati nel tentativo di riassumere e interpretare un certo numero di risultati di base utili nel contesto dell'elaborazione numerica dei segnali. Particolare attenzione è stata dedicata alle proprietà delle sequenze di correlazione e covarianza, allo spettro di potenza, e alle relazioni ingresso-uscita per sistemi discreti lineari invarianti alla traslazione. Nello scegliere tali argomenti siamo stati guidati soprattutto dalle esigenze poste dai successivi capitoli 9 e 11.

BIBLIOGRAFIA

1. W. B. Davenport, *Probability and Random Processes*, McGraw-Hill Book Company, New York, 1970.
2. A. W. Drake, *Fundamentals of Applied Probability Theory*, McGraw-Hill Book Company, New York, 1967.
3. W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. 1, 1968, Vol. 2, 1966, John Wiley & Sons, Inc., New York.
4. E. Parzen, *Modern Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1960.
5. W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill Book Company, New York, 1958.

6. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Company, New York, 1965.
7. E. Parzen, *Stochastic Processes*, Holden-Day, Inc., San Francisco, 1962.
8. G. E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day, Inc., San Francisco, 1970.
9. G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, Inc., San Francisco, 1968.

PROBLEMI

1. Dimostrare le seguenti proprietà delle medie:

(a) $E[x_n + y_m] = E[x_n] + E[y_m]$.

(b) $E[ax_n] = aE[x_n]$.

2. Si considerino due segnali aleatori scorrelati $x(n)$ e $y(n)$. Dimostrare che, se

$$w(n) = x(n) + y(n)$$

allora

$$m_w = m_x + m_y$$

e

$$\sigma_w^2 = \sigma_x^2 + \sigma_y^2$$

3. Per modellare gli effetti dell'arrotondamento e del troncamento nella realizzazione dei filtri numerici, rappresenteremo le variabili quantizzate come

$$y(n) = Q[x(n)] = x(n) + e(n)$$

dove $Q[\]$ indica arrotondamento o troncamento ed $e(n)$ è l'errore di quantizzazione. Sotto opportune ipotesi, è ragionevole ammettere che la sequenza $e(n)$ è una sequenza di rumore bianco, vale a dire

$$E[(e(n) - m_e)(e(n+m) - m_e)] = \sigma_e^2 \delta(m)$$

Dimostriamo nel cap. 9 che, nel caso dell'arrotondamento, la densità di probabilità del primo ordine è uniforme come mostrato nella fig. P8.3(a). Per il caso del troncamento la densità di probabilità del primo ordine è del tipo mostrato nella fig. P8.3(b).

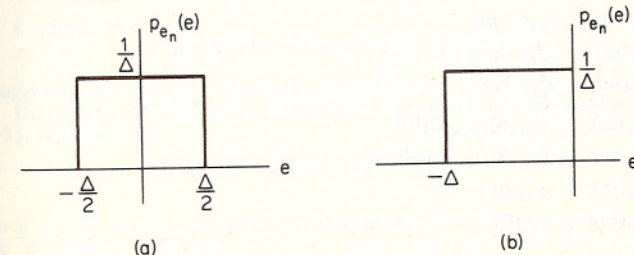


Fig. P8.3

- Trovare la media e la varianza per il rumor edovuto all'arrotondamento.
 - Trovare la media e la varianza per il rumore dovuto al troncamento.
4. Si indichi con $e(n)$ una sequenza di rumore bianco e con $s(n)$ una sequenza scorrelata con $e(n)$. Dimostrare che la sequenza

$$y(n) = s(n)e(n)$$

è bianca, cioè

$$E[y(n)y(n+m)] = A\delta(m)$$

dove A è una costante.

5. Si consideri un processo casuale per il quale le sequenze campione $x(n)$ sono della forma

$$x(n) = \cos(\omega_0 n + \theta)$$

dove θ è una variabile aleatoria a densità uniforme come in fig. P8.5. Si calcolino la media e la sequenza di autocorrelazione $\phi_{xx}(m, n)$. Il processo è stazionario in senso lato?

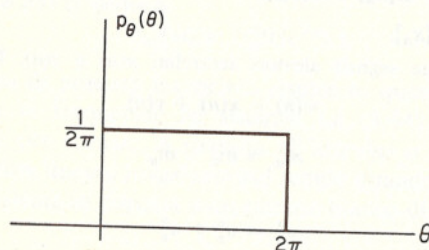


Fig. P8.5

6. Si considerino due processi casuali reali e stazionari, $\{x_n\}$ e $\{y_n\}$, con medie e varianze rispettivamente m_x ed m_y , σ_x^2 e σ_y^2 . Dimostrare che

$$\begin{aligned} (a) \quad \gamma_{xx}(m) &= \phi_{xx}(m) - m_x^2 \\ \gamma_{xy}(m) &= \phi_{xy}(m) - m_x m_y \end{aligned}$$

$$(b) \quad \phi_{xx}(0) = \text{media quadratica} \\ \gamma_{xx}(0) = \sigma_x^2$$

$$\begin{aligned} (c) \quad \phi_{xx}(m) &= \phi_{xx}(-m) \\ \gamma_{xx}(m) &= \gamma_{xx}(-m) \\ \phi_{xy}(m) &= \phi_{yx}^*(-m) \\ \gamma_{xy}(m) &= \gamma_{yx}^*(-m) \end{aligned}$$

$$\begin{aligned} (d) \quad |\phi_{xy}(m)| &\leq [\phi_{xx}(0)\phi_{yy}(0)]^{1/2} \\ |\gamma_{xy}(m)| &\leq [\gamma_{xx}(0)\gamma_{yy}(0)]^{1/2} \\ |\phi_{xx}(m)| &\leq \phi_{xx}(0) \\ |\gamma_{xx}(m)| &\leq \gamma_{xx}(0) \end{aligned}$$

Suggerimento: utilizzare la disuguaglianza

$$0 \leq E \left\{ \left(\frac{x_n}{(E[x_n^2])^{1/2}} - \frac{y_{n+m}}{(E[y_{n+m}^2])^{1/2}} \right)^2 \right\}$$

- (e) Se $y_n = x_{n-n_0}$, allora

$$\begin{aligned} \phi_{yy}(m) &= \phi_{xx}(m) \\ \gamma_{yy}(m) &= \gamma_{xx}(m) \end{aligned}$$

- (f) Siano $\Gamma_{xx}(z)$ e $\Gamma_{xy}(z)$ le trasformate z rispettivamente di $\gamma_{xx}(m)$ e $\gamma_{xy}(m)$. Dimostrare che:

$$(1) \quad \sigma_x^2 = \frac{1}{2\pi j} \oint_C \Gamma_{xx}(z) z^{-1} dz$$

$$\begin{aligned} (2) \quad \Gamma_{xx}(z) &= \Gamma_{xx}(1/z) \\ \Gamma_{xy}(z) &= \Gamma_{yx}^*(1/z^*) \end{aligned}$$

7. Si consideri un processo casuale *ergodico*, per il quale cioè le medie temporali coincidono con le medie d'insieme. Sia $x(n)$ una particolare sequenza campione del processo definito dall'insieme di variabili aleatorie $\{x_n\}$, $-\infty < n < +\infty$, dove $p_x(x)$ è la densità di probabilità del primo ordine per tutte le variabili aleatorie x_n .

(a) Considerare la media temporale della funzione $u(a - x_n)$, cioè $\langle u(a - x_n) \rangle$. (La funzione $u(\cdot)$ è il gradino unitario). Esprimere a parole il significato di questa media temporale.

(b) Considerare la media d'insieme della funzione $u(a - x_n)$. Che cosa vale $E[u(a - x_n)]$ in termini di $p_x(x)$?

(c) Sono coerenti la risposta data in (b) e l'interpretazione data in (a)? In altri termini, è ragionevole che

$$E[u(a - x_n)] = \langle u(a - x_n) \rangle?$$

8. Abbiamo osservato che la sequenza di autocorrelazione serve come indicazione della rapidità di variazione di un segnale casuale. Questa idea si può formalizzare considerando la media quadratica della variazione tra campione e campione del segnale reale, definita da

$$E[(x(n+1) - x(n))^2]$$

Si consideri un segnale $x(n)$ a media zero e con spettro di potenza a banda limitata tale che

$$P_{xx}(\omega) = 0, \quad \omega_c < |\omega| \leq \pi$$

- (a) Dimostrare che

$$E[(x(n+1) - x(n))^2] = 2[\phi_{xx}(0) - \phi_{xx}(1)]$$

- (b) Dimostrare che

$$[\phi_{xx}(0) - \phi_{xx}(1)] \leq \frac{\omega_c^2}{2} \phi_{xx}(0)$$

e quindi che

$$E[(x(n+1) - x(n))^2] \leq \omega_c^2 E[x^2(n)]$$

[Suggerimento: usare il fatto che $\sin^2(\omega/2) \leq \omega^2/4$ per $0 \leq \omega \leq \omega_c$].

Una disuguaglianza molto importante, nota come disuguaglianza di Chebyshev [1-5], stabilisce che

$$\text{Probabilità } [|x(n) - E[x(n)]| \geq \epsilon] \leq \frac{\sigma_x^2}{\epsilon^2}$$

per $\epsilon > 0$. Si osservi che questa disuguaglianza consente di esprimere l'aspettativa che un segnale differisca dalla sua media, nel senso che più grande è il suo valore quadratico medio, più elevata è la probabilità che un valore particolare della sequenza differisca dal valor medio di una quantità maggiore di ϵ .

- (c) Usare il risultato di (b) e la disuguaglianza di Chebyshev per dimostrare che

$$\text{Probabilità } [|x(n+1) - x(n)| > \epsilon] \leq \frac{\omega_c^2 E[x^2(n)]}{\epsilon^2}$$

per $\epsilon > 0$. Interpretare questo risultato alla luce delle proprie conoscenze sulla relazione che esiste fra le variazioni nel tempo di un segnale e la sua larghezza di banda.

9. Sia $x(n)$ un rumore con media zero e varianza σ_x^2 . Sia $y(n)$ l'uscita corrispondente quando $x(n)$ è in ingresso a un sistema lineare invariante alla traslazione con risposta all'impulso $h(n)$.

(a) Dimostrare che

$$E[x(n)y(n)] = h(0)\sigma_x^2$$

(b) Dimostrare che

$$\sigma_y^2 = \sigma_x^2 \sum_{n=-\infty}^{\infty} h^2(n)$$

10. Dimostrare che la trasformata z di

$$v(n) = \sum_{k=-\infty}^{\infty} h(k)h(k+n)$$

è

$$V(z) = H(z)H(z^{-1})$$

11. Un sistema ideale che effettua la trasformata di Hilbert e che ha pertanto la risposta all'impulso

$$h(n) = \begin{cases} \frac{2 \sin^2(\pi n/2)}{\pi n} & n \neq 0 \\ 0 & n = 0 \end{cases}$$

è eccitato da un segnale casuale a tempo discreto $x_r(n)$ come indicato in fig. P8.11.



Fig. P8.11

- (a) Trovare una espressione per la sequenza di autocorrelazione $\phi_{x_i x_i}(m)$.
 (b) Trovare una espressione per la sequenza di correlazione incrociata $\phi_{x_r x_i}(m)$. Mostrare che in questo caso $\phi_{x_r x_i}(m) = -\phi_{x_r x_i}(-m)$.
 (c) Trovare la sequenza di autocorrelazione del segnale analitico complesso $x(n) = x_r(n) + jx_i(n)$.
 (d) Trovare lo spettro di potenza $P_{xx}(\omega)$ del segnale analitico complesso di (c).
 12. Si consideri la rete numerica lineare invariante alla traslazione indicata in fig. P8.12. Siano $h_1(n)$ e $h_2(n)$ le risposte all'impulso tra rispettivamente i nodi 1 e 2 e l'uscita. Dimostrare che se $x_1(n)$ e $x_2(n)$ sono scorrelati, sono scorrelati anche le corrispondenti uscite.

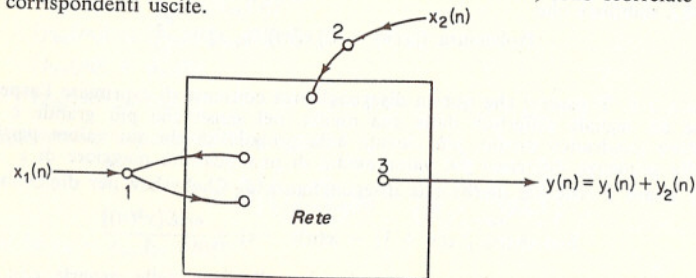


Fig. P8.12

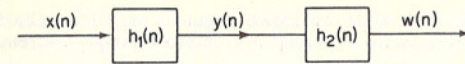


Fig. P8.13

13. Sia $x(n)$ una sequenza casuale bianca con media zero e varianza σ_x^2 . Si ponga $x(n)$ all'ingresso della cascata di due sistemi lineari a tempo discreto e invarianti alla traslazione, come indicato in fig. P8.13.

(a) È $\sigma_y^2 = \sigma_x^2 \sum_{k=0}^{\infty} h_1^2(k)$?

(b) È $\sigma_w^2 = \sigma_y^2 \sum_{k=0}^{\infty} h_2^2(k)$?

- (c) Siano $h_1(n) = a^n u(n)$ e $h_2(n) = b^n u(n)$. Determinare la risposta all'impulso dell'intero sistema di fig. P8.13, e da questa ricavare σ_w^2 . Se la risposta alla domanda (b) era positiva, c'è coerenza con la soluzione trovata in (c)?

14. Poiché in molte applicazioni i segnali casuali a tempo discreto nascono dal campionamento di segnali casuali a tempo continuo, tratteremo in questo problema la derivazione del teorema del campionamento per segnali casuali. Si consideri un processo casuale a tempo continuo definito dalle variabili aleatorie $\{x_s(t)\}$, dove t è una variabile continua. La funzione di autocovarianza, assumendo che la media sia nulla, è definita come

$$\gamma_{x_s x_s}(\tau) = E[x_s(t)x_s^*(t + \tau)]$$

e lo spettro di potenza è

$$P_{x_s x_s}(\Omega) = \int_{-\infty}^{\infty} \gamma_{x_s x_s}(\tau) e^{-j\Omega\tau} d\tau$$

Un processo casuale a tempo discreto ottenuto mediante campionamento periodico è definito dall'insieme di variabili aleatorie $\{x(n)\}$, dove $x(n) = x_s(nT)$ e T è il periodo di campionamento.

- (a) Quale è la relazione fra $\gamma_{xx}(n)$ e $\gamma_{x_s x_s}(\tau)$?
 (b) Esprimere lo spettro di densità di potenza del processo a tempo discreto per mezzo dello spettro di densità di potenza del processo a tempo continuo.
 (c) Sotto quali condizioni lo spettro del processo a tempo discreto è una fedele rappresentazione dello spettro del processo a tempo continuo?
 15. Si consideri un processo casuale a tempo continuo $\{x_s(t)\}$ con spettro di potenza limitato in banda come mostrato in fig. P8.15-1. Supponiamo di campionare $\{x_s(t)\}$ in modo da ottenere il processo casuale a tempo discreto $\{x(n) = x_s(nT)\}$.

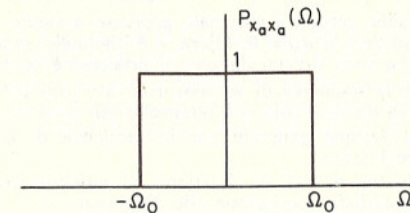


Fig. P8.15-1

- (a) Quale è la sequenza di autocovarianza del processo casuale a tempo discreto?
 (b) Per lo spettro di potenza analogico dato sopra, come deve essere scelto T perché il processo a tempo discreto sia bianco?

- (c) Se lo spettro di potenza analogico è quello mostrato nella fig. P8.15-2, come dovrebbe essere scelto T perché il processo a tempo discreto sia bianco?

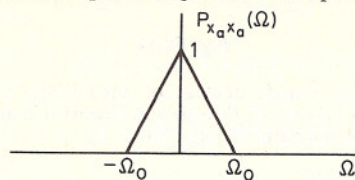


Fig. P8.15-2

- (d) Quale è la condizione generale da imporre al processo analogico e al periodo di campionamento in modo che il processo a tempo discreto risulti bianco?
16. Spesso è conveniente assumere che un processo casuale stazionario in senso lato nasce come risultato dell'eccitazione di un sistema lineare per mezzo di un rumore bianco. Tali processi si chiamano *processi lineari*. Si consideri un sistema lineare stabile come mostrato in fig. P8.16, dove $x(n)$ è un rumore bianco con media nulla e varianza σ_x^2 .

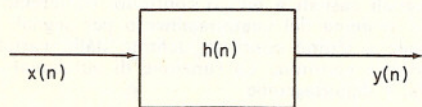


Fig. P8.16

- (a) Esprimere l'autocovarianza di $y(n)$ in termini della risposta all'impulso del sistema.
- (b) Usare il risultato di (a) per esprimere lo spettro di potenza di $y(n)$ in termini della risposta in frequenza del sistema.
- È di particolare interesse il caso in cui $H(z)$ è una funzione razionale della forma

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

nel qual caso il processo $y(n)$ è legato al processo rumore bianco $x(n)$ dall'equazione alle differenze

$$y(n) = \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^M b_k x(n-k)$$

Se tutti gli a_k sono zero, y è chiamato *processo a media mobile*. Se tutti i b_k , eccetto b_0 , sono zero e $a_N \neq 0$, allora y è chiamato *processo autoregressivo*. Se sia a_N che b_M sono diversi da zero il processo è *misto*.

- (c) Dimostrare che la sequenza di autocovarianza $\gamma_{yy}(m)$ di un processo a media mobile è diversa da zero solo nell'intervallo $|m| \leq M$.
- (d) Trovare un'espressione generale per la sequenza di autocovarianza di un processo autoregressivo.
- (e) Dimostrare che, se $b_0 = 1$, la funzione di autocovarianza di un processo autoregressivo soddisfa l'equazione alle differenze

$$\gamma_{yy}(0) = \sum_{k=1}^N a_k \gamma_{yy}(k) + \sigma_x^2,$$

$$\gamma_{yy}(m) = \sum_{k=1}^N a_k \gamma_{yy}(m-k), \quad m \geq 1$$

- (f) Usare il risultato di (e) e la simmetria di $\gamma_{yy}(m)$ per dimostrare la validità del seguente insieme di identità

$$\sum_{k=1}^N a_k \gamma_{yy}(|m-k|) = \gamma_{yy}(m), \quad m = 1, 2, \dots, N$$

Pertanto, dati i primi $N+1$ valori della covarianza, si possono sempre ricavare in maniera univoca i valori di a_k e σ_x^2 che caratterizzano il processo y .

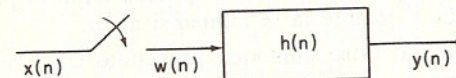
17. Si consideri il processo lineare *misto* y che soddisfa l'equazione alle differenze

$$y(n) = \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^M b_k x(n-k), \quad b_0 = 1$$

Talora ha interesse « imbiancare » lo spettro di $y(n)$ trasformando $y(n)$ per mezzo di un filtro lineare; si desidera, cioè, trovare un sistema tale che l'uscita abbia spettro di potenza piatto se in ingresso c'è $y(n)$.

Supponiamo di conoscere la funzione di autocovarianza $\gamma_{yy}(m)$ e la sua trasformata z , $\Gamma_{yy}(z)$, ma non gli a_k e i b_k .

- (a) Discutere una procedura o trovare una funzione di trasferimento $H_w(z)$ del filtro imbiancante.
- (b) Il filtro imbiancante è unico?
- (c) Come intervengono le trasformate di Hilbert (del cap. 7) in questo problema?
18. Talora si è interessati al comportamento statistico di un sistema lineare invariante alla traslazione quando in ingresso viene improvvisamente applicato un



(Interruttore chiuso in $n=0$)

Fig. P8.18

segnale casuale. Una tale situazione è mostrata nella fig. P8.18. Sia $x(n)$ un rumore bianco. L'ingresso al sistema

$$w(n) = \begin{cases} x(n), & n \geq 0 \\ 0, & n < 0 \end{cases}$$

è un processo non stazionario, come anche l'uscita $y(n)$.

- (a) Derivare un'espressione della media dell'uscita in funzione della media dell'ingresso.
- (b) Derivare un'espressione per la sequenza di autocorrelazione $\phi_{yy}(n_1, n_2)$ dell'uscita.
- (c) Mostrare che per n grande le formule derivate in (a) e (b) approssimano i risultati per ingressi stazionari.
- (d) Si assuma che $h(n) = a^n u(n)$. Trovare la media e il valore quadratico medio dell'uscita in termini della media e del valore quadratico medio dell'ingresso. Dare una rappresentazione grafica della dipendenza di questi parametri da n .

9. EFFETTI DELLA LUNGHEZZA FINITA DEI REGISTRI NELLA ELABORAZIONE NUMERICA DEI SEGNALI

9.0 INTRODUZIONE

Gli algoritmi di elaborazione numerica dei segnali come quelli per il filtraggio lineare e per le trasformazioni di Fourier discrete si realizzano sia con circuiti specializzati (« special purpose ») sia con programmi per calcolatori di uso generale (« general purpose »). In entrambi i casi i valori delle sequenze e i coefficienti vengono memorizzati con formato binario in registri di lunghezza finita. Questa lunghezza finita di parola costituisce una limitazione che si manifesta in numerosi modi.

I parametri di un filtro numerico progettato con una delle tecniche del cap. 5 si ottengono generalmente con elevata precisione. Quando tuttavia questi parametri vengono quantizzati, la risposta in frequenza del filtro numerico risultante può differire apprezzabilmente da quella originalmente progettata. In effetti il filtro quantizzato può non soddisfare quelle stesse specifiche che sono invece soddisfatte dal filtro non quantizzato. Nel cap. 4 si è visto che la sensibilità della risposta del filtro ad errori nei parametri dipende dalla struttura scelta per realizzare il filtro. I relativi problemi di scelta della struttura del filtro e il progetto diretto di filtri con coefficienti quantizzati sono tuttora importanti argomenti di ricerca.

Quando una sequenza da elaborare deriva dal campionamento di un segnale analogico a banda limitata, la limitazione dovuta alla lunghezza finita di parola comporta che nel processo di conversione da analogico a numerico siano disponibili soltanto un numero finito di possibili valori per ogni campione. In altri termini, i campioni del segnale devono essere quantizzati per adattarsi alla lunghezza finita dei registri. Vedremo nel presente capitolo che questo effetto può spesso essere trattato in termini di un segnale di rumore aggiuntivo.

Anche quando si parte con dati rappresentabili con lunghezza finita di parola, il risultato dell'elaborazione porta generalmente a numeri che richiedono ulteriori bit per la loro rappresentazione. Per esempio, moltiplicando tra loro un campione e un coefficiente rappresentati con b bit si ha un prodotto che è lungo $2b$ bit. Se allora, in una realizzazione ricorsiva di un filtro numerico, non si quantizza il risultato delle operazioni aritme-

tiche, il numero dei bit crescerebbe indefinitamente, in quanto dopo la prima iterazione sarebbero necessari $2b$ bit, dopo la seconda $3b$ bit, ecc. Analogamente, in un algoritmo di FFT, se i coefficienti sono numeri a b bit, la rappresentazione precisa del risultato di ogni stadio del calcolo della FFT richiede b bit in più rispetto allo stadio precedente. L'effetto della quantizzazione in simili casi dipende dal fatto che l'aritmetica usata sia a virgola fissa o a virgola mobile, o che i numeri in virgola fissa rappresentino frazioni o interi, o che la quantizzazione sia fatta mediante troncamento o arrotondamento. In questo capitolo tratteremo separatamente i casi dell'aritmetica in virgola fissa e in virgola mobile. Per l'aritmetica in virgola fissa è naturale, in un contesto di elaborazione di segnali, considerare un registro come rappresentante una frazione a virgola fissa. In questo modo il prodotto di due numeri rimane una frazione e la lunghezza limitata di registro può essere mantenuta troncando o arrotondando i bit meno significativi. Con questo tipo di rappresentazione il risultato della somma di frazioni a virgola fissa non richiede di essere troncato o arrotondato. Tuttavia il modulo della somma risultante può essere più grande di uno. Questo effetto è comunemente chiamato *saturazione* (« overflow ») e lo si può fronteggiare richiedendo che i dati di ingresso siano sufficientemente piccoli in modo da evitare la possibilità di saturazione. Nel caso dell'aritmetica a virgola mobile, grazie all'ampio campo di numeri rappresentabili, si possono generalmente evitare simili considerazioni sulla dimensione dei numeri da trattare; la quantizzazione interviene tuttavia sia per la moltiplicazione che per l'addizione.

Nella discussione che segue cominceremo col passare in rassegna le rappresentazioni in virgola fissa e in virgola mobile dei numeri binari e le rappresentazioni dei numeri negativi in complemento a uno, in complemento a due e in modulo e segno. Discuteremo poi la relazione fra la rappresentazione binaria e il troncamento o l'arrotondamento. Troncamento o arrotondare il risultato di un'operazione aritmetica corrisponde a inserire una non linearità nel filtro in oggetto. Per filtri e ingressi abbastanza semplici è possibile analizzare l'effetto di questa non linearità in termini di quello che viene chiamato *comportamento di ciclo limite del filtro*. Per filtri e ingressi alquanto complessi questa analisi è invece molto difficile da svolgersi. In tal caso è spesso utile effettuare un'analisi approssimata rappresentando l'effetto del troncamento o dell'arrotondamento in termini di un segnale di errore aggiuntivo che verrà chiamato *rumore di arrotondamento*. Il filtro si considera allora lineare, ma l'uscita ha una componente di rumore che risulta dall'effetto dell'arrotondamento o troncamento. Le proprietà medie di questo rumore possono essere analizzate usando le tecniche del cap. 8. Discuteremo un'analisi di rumore di questo tipo per realizzazioni in virgola fissa e in virgola mobile di filtri numerici ed anche per la trasformata di Fourier veloce.

9.1 EFFETTO DELLA RAPPRESENTAZIONE DEI NUMERI SULLA QUANTIZZAZIONE

9.1.1 Numeri binari in virgola fissa e in virgola mobile

Nelle realizzazioni numeriche sia « hardware » che « software » si usa generalmente una rappresentazione dei numeri in base due, e cioè una rappresentazione binaria [1]. Pertanto un numero è rappresentato mediante una sequenza di cifre binarie (*bit*) che sono o zero o uno. Proprio come un numero decimale è rappresentato da una sequenza di cifre decimali con una virgola che divide la parte intera dalla parte frazionaria, così la sequenza di cifre binarie è divisa da una virgola binaria in quelle che rappresentano la parte intera del numero e in quelle che ne rappresentano la parte frazionaria. Pertanto, se Δ sta ad indicare la posizione della virgola binaria, il numero binario 1001 Δ 0110 ha il valore decimale di

$$(1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0) + (0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 0 \cdot 2^{-4}),$$

o 9.375.

Le aritmetiche usate nelle realizzazioni « hardware » o « software » si differenziano essenzialmente per la posizione della virgola binaria. Per l'aritmetica in virgola fissa, la realizzazione è basata sull'assunto che la posizione della virgola binaria è fissa. Il modo in cui si effettua la somma non dipenderà, per l'aritmetica in virgola fissa, dalla posizione della virgola binaria, purché essa sia la medesima per ogni registro.

Per la moltiplicazione è generalmente molto conveniente assumere o che tutti i numeri sono interi, o che essi sono tutti delle frazioni (proprie), in quanto il prodotto di interi è un intero e il prodotto di frazioni è ancora una frazione. Nelle applicazioni proprie del filtraggio numerico è solitamente necessario approssimare il prodotto di due numeri a b bit, che avrebbe $2b$ bit, con un risultato ancora a b bit. Nell'aritmetica intera ciò è difficile. Con l'aritmetica delle frazioni, d'altra parte, ciò può essere fatto troncando o arrotondando il numero ai b bit più significativi. Per la moltiplicazione con frazioni non può mai verificarsi la saturazione in quanto il prodotto di due frazioni è ancora una frazione. Per esempio, se moltiplichiamo le due frazioni a quattro bit $\Delta 1001$ e $\Delta 0011$, il prodotto a otto bit $\Delta 00011011$ può essere approssimato con $\Delta 0001$ (troncamento) o $\Delta 0010$ (arrotondamento).

Se sommiamo due frazioni a virgola fissa, può verificarsi la saturazione. Per esempio, la somma delle due frazioni a quattro bit $\Delta 1101$ e $\Delta 1000$ è $1\Delta 0101$. Ne segue che la somma non può essere contenuta in un registro a quattro bit. Questa limitazione sulla dinamica dei numeri rappresentabili può essere nella sostanza rimossa usando una rappresentazione in virgola mobile. Nella più comune rappresentazione a virgola mo-

bile, un numero positivo F è rappresentato come $F = 2^c \cdot M$, dove M , la mantissa, è una frazione, tale che

$$\frac{1}{2} \leq M < 1$$

e c , la caratteristica, può essere sia positiva che negativa. Quando M è nel suddetto campo di valori, la rappresentazione in virgola mobile si dice normalizzata. Il prodotto di due numeri in virgola mobile si effettua moltiplicando le mantisse come frazioni in virgola fissa e sommando le caratteristiche. Poiché il prodotto delle mantisse sarà un numero fra $1/4$ e 1 , potrà rendersi necessaria una normalizzazione della mantissa e il corrispondente aggiustamento della caratteristica. Vale a dire che, se $M < 1/2$, M viene traslato a sinistra di un bit e la caratteristica si incrementa di uno¹.

La somma di due numeri in virgola mobile si effettua traslando verso destra i bit della mantissa del numero più piccolo finché le caratteristiche dei due numeri sono uguali e sommando quindi le mantisse come illustrato nell'esempio che segue.

ESEMPIO. Si consideri la somma di F_1 ed F_2 con $F_1 = 4$ ed $F_2 = 5/4$. Usando la notazione in virgola mobile, è $F_1 = 2^{c_1}M_1$, ed $F_2 = 2^{c_2}M_2$, con

$$c_1 = 11_{\Delta} \quad (= 3 \text{ in decimale})$$

$$M_1 = \Delta 10000 \quad (= \frac{1}{2} \text{ in decimale})$$

$$c_2 = 01_{\Delta} \quad (= 1 \text{ in decimale})$$

$$M_2 = \Delta 10100 \quad (= \frac{5}{8} \text{ in decimale})$$

Per effettuare la somma, c_2 deve diventare uguale a c_1 , ed M_2 deve cambiare di conseguenza. Pertanto si comincia col cambiare la rappresentazione di F_2 in $F_2 = 2^{c_2} \cdot M_2$, con

$$\hat{c}_2 = 11_{\Delta}$$

$$\hat{M}_2 = \Delta 00101$$

per cui le mantisse possono ora essere sommate. La somma risultante è $F = 2^c \cdot M$, con $c = 11$ ed $M = \Delta 10101$. In questo caso la somma di M_1 ed M_2 è una frazione tra $1/2$ ed 1 , per cui non è necessario un ulteriore aggiustamento di c . In un caso più generale, la somma delle mantisse potrebbe non trovarsi nel corretto campo di valori e di conseguenza c verrebbe modificato per portare la mantissa nel campo corretto.

Da questo esempio dovrebbe risultare chiaro che con l'aritmetica in virgola mobile la mantissa può eccedere le dimensioni del registro sia per l'addizione che per la moltiplicazione e va pertanto o troncata o arrotondata, laddove, per l'aritmetica in virgola fissa, ciò è necessario solo nel caso della moltiplicazione. D'altra parte, se nel caso della virgola fissa il risultato di una somma eccede la lunghezza del registro, il troncamento o l'arrotondamento non sono di alcun aiuto e si ha saturazione. Pertanto la virgola mobile, se da un lato introduce errori di arrotondamento sia per la

¹ Traslare la mantissa di un posto a destra corrisponde a dividere per 2 mentre traslarla di un posto a sinistra corrisponde a moltiplicare per 2. Pertanto la caratteristica si incrementa quando la mantissa è traslata a destra e diminuisce quando la mantissa è traslata a sinistra.

somma che per la moltiplicazione, d'altro lato consente di trattare numeri con una dinamica di valori molto più ampia che nel caso della virgola fissa. Si devono considerare entrambi questi effetti quando si vogliono confrontare realizzazioni di filtri numerici in virgola fissa e in virgola mobile.

9.1.2 Rappresentazione dei numeri negativi

Nella elaborazione numerica dei segnali, come nella maggior parte degli algoritmi numerici, è necessario trattare numeri con segno. Esistono usualmente tre metodi per rappresentare numeri negativi a virgola fissa. Il primo, e forse il più noto, è quello del *modulo e segno*. In questa rappresentazione, il modulo (che è ovviamente positivo) è rappresentato come un numero binario e il segno è rappresentato mediante la cifra binaria di testa: 0 indica un numero positivo e 1 indica un numero negativo (o viceversa). Per esempio, in modulo e segno $0_{\Delta}0011$ rappresenta $3/16$ mentre $1_{\Delta}0011$ rappresenta $-3/16$.

Le altre due più comuni rappresentazioni dei numeri negativi si chiamano rappresentazioni in complemento a due e in complemento a uno. La *rappresentazione in complemento a due* può essere vista in termini di una interpretazione di tutti i numeri nel registro come numeri positivi. Con un totale di $(b + 1)$ bit (uno alla sinistra e b alla destra della virgola binaria), questi numeri, se interpretati come positivi, vanno da zero a $2 - 2^{-b}$. Metà di tali numeri sono usati per rappresentare le frazioni positive e metà quelle negative. Più precisamente, le frazioni positive sono rappresentate come in modulo e segno e le frazioni negative sono rappresentate sottraendo l'ampiezza da 2.0 .

ESEMPIO. La frazione in modulo e segno $0_{\Delta}0110$ è, in complemento a due, il numero $0_{\Delta}0110$. Si consideri, tuttavia, il numero in modulo e segno $1_{\Delta}0110$. Il modulo è $0_{\Delta}0110$ (vale a dire $3/8$), che se sottratto da $10_{\Delta}0000$ (cioè 2) dà luogo a $1_{\Delta}1010$ (cioè $1 + 5/8$). Pertanto $-3/8$ è rappresentato da $1_{\Delta}1010$.

La *rappresentazione in complemento a uno* dei numeri negativi è simile a quella in complemento a due. Le frazioni positive sono ancora rappresentate come in modulo e segno. Le frazioni negative sono rappresentate sottraendone il modulo dal numero più grande che entra nel registro, cioè quello con tutti i bit uguali a uno. Pertanto con $(b + 1)$ bit come prima (b bit alla destra e un bit alla sinistra della virgola binaria), le frazioni negative sono rappresentate sottraendone il modulo dal numero $2 - 2^{-b}$.

ESEMPIO. Si assuma $b = 4$. La frazione positiva $+3/8$ è rappresentata in complemento a uno come $0_{\Delta}0110$. La frazione negativa $(-3/8)$ ha modulo $(3/8)$. Pertanto la rappresentazione in complemento a uno si ottiene sottraendo $0_{\Delta}0110$ ($3/8$) da $1_{\Delta}1111$ ($2 - 2^{-4}$) dando luogo a $1_{\Delta}1001$. Di conseguenza la frazione negativa $(-3/8)$ è rappresentata da $1_{\Delta}1001$ in un registro in complemento a uno.

Tab. 9.1

Interpretazione

Numero binario	Modulo e segno	Complemento a due	Complemento a uno
$0_{\Delta}111$	$7/8$	$7/8$	$7/8$
$0_{\Delta}110$	$6/8$	$6/8$	$6/8$
$0_{\Delta}101$	$5/8$	$5/8$	$5/8$
$0_{\Delta}100$	$4/8$	$4/8$	$4/8$
$0_{\Delta}011$	$3/8$	$3/8$	$3/8$
$0_{\Delta}010$	$2/8$	$2/8$	$2/8$
$0_{\Delta}001$	$1/8$	$1/8$	$1/8$
$0_{\Delta}000$	0	0	0
$1_{\Delta}000$	-0	-1	-7/8
$1_{\Delta}001$	$-1/8$	$-7/8$	$-6/8$
$1_{\Delta}010$	$-2/8$	$-6/8$	$-5/8$
$1_{\Delta}011$	$-3/8$	$-5/8$	$-4/8$
$1_{\Delta}100$	$-4/8$	$-4/8$	$-3/8$
$1_{\Delta}101$	$-5/8$	$-3/8$	$-2/8$
$1_{\Delta}110$	$-6/8$	$-2/8$	$-1/8$
$1_{\Delta}111$	$-7/8$	$-1/8$	-0

La tabella 9.1 mostra un confronto delle tre rappresentazioni dei numeri per una lunghezza di parola di quattro bit. È utile notare che in tutte e tre le rappresentazioni il bit di testa è zero per una frazione positiva e uno per una frazione negativa (v. il probl. 1(a) alla fine di questo capitolo). Per questa ragione il bit di testa è chiamato il *bit segno*. Per la rappresentazione in modulo e segno, il cambiare un numero in segno ma non in modulo influisce solo sul bit di testa. Per le rappresentazioni in complemento a uno e in complemento a due, il cambiamento di segno di un numero influisce su tutti i bit. In particolare, si può mostrare che cambiare il segno di un numero in complemento a uno corrisponde a cambiare tutti i suoi bit (v. il probl. 1(b) di questo capitolo); cambiare il segno di un numero in complemento a due corrisponde invece a cambiare tutti i suoi bit e aggiungere 2^{-b} , ignorando ogni eventuale saturazione conseguente all'addizione (v. il probl. 1(c) alla fine di questo capitolo). Si noti anche che sia $+0$ che -0 sono rappresentati nei sistemi in modulo e segno e in complemento a uno; nel sistema in complemento a due è rappresentato -1 , ma non $+1$. Ciascuna rappresentazione ha i suoi vantaggi e svantaggi; la scelta della rappresentazione dei numeri negativi dipende essenzialmente dalla realizzazione « hardware » o « software » delle operazioni aritmetiche, quali l'addizione, la sottrazione e la moltiplicazione.

Per la rappresentazione dei numeri negativi in virgola mobile sono state usate numerose convenzioni. In questo capitolo assumeremo che il segno del numero è associato alla mantissa, così che la mantissa è una frazione con segno. La rappresentazione di quest'ultima può ovviamente essere fatta in modulo e segno, in complemento a uno o in complemento a due.

9.1.3 Effetto del troncamento o arrotondamento

Faremo l'ipotesi che sia i numeri in virgola fissa che le mantisse dei numeri in virgola mobile siano rappresentati come frazioni binarie a $(b + 1)$ bit, con la virgola binaria posta immediatamente a destra del bit di ordine più alto. Questa convenzione non comporta perdita di generalità, e alla sua convenienza si è già fatto cenno sopra. Il valore numerico (per numeri positivi) di un uno nel bit meno significativo è 2^{-b} . Questa quantità verrà chiamata *ampiezza di quantizzazione* in quanto i numeri sono quantizzati a intervalli di 2^{-b} .

Come indicato in precedenza, l'effetto del troncamento o arrotondamento dipende dall'usare un'aritmetica in virgola fissa o virgola mobile e da come sono rappresentati i numeri negativi. Consideriamo per primo l'effetto del troncamento e arrotondamento nel caso della virgola fissa. La rappresentazione dei numeri positivi è la stessa per i metodi modulo e segno, complemento a uno e complemento a due, e pertanto identico sarà l'effetto del troncamento o arrotondamento. Indichiamo con b_1 il numero di bit alla destra della virgola binaria prima del troncamento e con b il numero di bit alla destra della virgola binaria dopo il troncamento, con, ovviamente, $b < b_1$. L'effetto del troncamento è quello di scartare i $(b_1 - b)$ bit meno significativi, e di conseguenza l'ampiezza del numero dopo il troncamento è minore o uguale all'ampiezza prima del troncamento.

Indichiamo con x e $Q(x)$ il numero rispettivamente prima e dopo il troncamento. L'errore di troncamento è pertanto

$$E_T = Q[x] - x$$

Questo errore per numeri positivi sarà negativo o nullo. Il massimo errore si verifica quando tutti i bit scartati sono uno, nel qual caso il troncamento riduce il valore del registro di $(2^{-b} - 2^{-b_1})$. Pertanto, per il troncamento di numeri positivi, è

$$-(2^{-b} - 2^{-b_1}) \leq E_T \leq 0 \quad (9.1)$$

Per numeri negativi l'effetto del troncamento dipende dalla rappresentazione usata (in modulo e segno, in complemento a due o complemento a uno). Considereremo pertanto separatamente ciascuno dei tre casi.

Con la rappresentazione in modulo e segno l'effetto del troncamento, come già accennato sopra, è quello di ridurre il valore assoluto del numero. Pertanto un numero negativo diventa più piccolo in modulo, e quindi E_T , cioè il valore dopo il troncamento meno il valore prima del troncamento, è positivo. Per i numeri negativi rappresentati in modulo e segno si ha dunque

$$0 \leq E_T \leq (2^{-b} - 2^{-b_1}) \quad (9.2)$$

Per un numero negativo in complemento a due costituito dalla sequenza di bit $1_a a_1 a_2 \dots a_{b_1}$, il modulo è dato da

$$A_1 = 2.0 - x_1$$

dove

$$x_1 = 1 + \sum_{i=1}^{b_1} a_i 2^{-i}$$

Il troncamento a b bit produce la sequenza di bit $1_a a_1 a_2 \dots a_b$, dove ora il modulo è

$$A_2 = 2.0 - x_2$$

con

$$x_2 = 1 + \sum_{i=1}^b a_i 2^{-i}$$

Il cambiamento in valore assoluto è

$$\Delta A = A_2 - A_1 = \sum_{i=b+1}^{b_1} a_i 2^{-i}$$

e si vede facilmente che

$$0 \leq \Delta A \leq 2^{-b} - 2^{-b_1}$$

Pertanto l'effetto del troncamento per numeri negativi in complemento a due è quello di *aumentare* il modulo del numero negativo; l'errore di troncamento è quindi negativo e si ha

$$-(2^{-b} - 2^{-b_1}) \leq E_T \leq 0 \quad (9.3)$$

Per un numero negativo in complemento a uno costituito dalla sequenza di bit $1_a a_1 a_2 \dots a_{b_1}$, il modulo è dato da $A_1 = 2.0 - 2^{-b_1} - x_1$, e il troncamento a b bit dà luogo al modulo $A_2 = 2.0 - 2^{-b} - x_2$, dove x_1 e x_2 sono da interpretare nel senso definito sopra. Il cambiamento in valore assoluto è

$$\Delta A = A_2 - A_1 = \sum_{i=b+1}^{b_1} a_i 2^{-i} - (2^{-b} - 2^{-b_1})$$

e quindi

$$-(2^{-b} - 2^{-b_1}) \leq \Delta A \leq 0$$

Pertanto l'effetto del troncamento per numeri negativi in complemento a uno è di *diminuire* il modulo del numero negativo; l'errore di troncamento è positivo e soddisfa la disequaglianza

$$0 \leq E_T \leq (2^{-b} - 2^{-b_1}) \quad (9.4)$$

Si osservi che per i numeri in complemento a due il campo di valori per l'errore è lo stesso per numeri positivi e negativi, mentre, nel caso del complemento a uno e del modulo e segno, il segno dell'errore dipende dal segno del numero che viene troncato.

Come alternativa al troncamento, i numeri possono essere arrotondati, sempre per rispettare la lunghezza finita dei registri. Indichiamo ancora con b il numero di bit alla destra della virgola binaria dopo l'arrotondamento. Dopo l'arrotondamento i valori sono quantizzati a intervalli di 2^{-b} , e, cioè, la più piccola differenza non nulla fra due numeri è 2^{-b} .

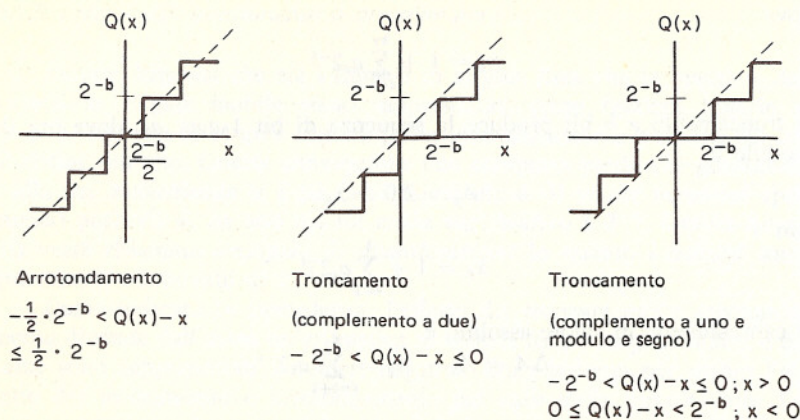


Fig. 9.1 Relazioni non lineari rappresentanti arrotondamento e troncamento

Arrotondare corrisponde a scegliere il livello di quantizzazione più vicino². Ne segue che l'errore massimo ha un valore assoluto di $0.5 \cdot 2^{-b}$, e cioè l'errore di arrotondamento E_R ha il campo di valori³

$$-\frac{1}{2}(2^{-b} - 2^{-b_1}) < E_R \leq \frac{1}{2}(2^{-b} - 2^{-b_1}) \quad (9.5)$$

Poiché l'arrotondamento è basato sul modulo del numero, l'errore è indipendente dal modo in cui si rappresentano i numeri negativi. Generalmente è ragionevole assumere che $2^{-b_1} \ll 2^{-b}$ e di conseguenza che il termine 2^{-b_1} può essere trascurato nelle disuguaglianze di sopra. Con questa approssimazione gli errori di troncamento e arrotondamento si possono riassumere come segue:

Troncamento:

$$-2^{-b} < E_T \leq 0 \quad \begin{array}{l} \text{Numeri positivi} \\ \text{e numeri negativi} \\ \text{in complemento} \\ \text{a due} \end{array} \quad (9.6a)$$

$$0 \leq E_T < 2^{-b} \quad \begin{array}{l} \text{Numeri negativi} \\ \text{in ampiezza e} \\ \text{segno e in com-} \\ \text{plemento a uno} \end{array} \quad (9.6b)$$

Arrotondamento:

$$-\frac{1}{2} \cdot 2^{-b} < E_R \leq \frac{1}{2} \cdot 2^{-b} \quad (9.6c)$$

Questi errori di troncamento e arrotondamento sono anche riassunti nella fig. 9.1.

Nell'aritmetica a virgola mobile, il troncamento o arrotondamento influisce solo sulla mantissa. Pertanto, nella rappresentazione a virgola

² È ovviamente possibile che il numero da arrotondare si trovi esattamente a metà strada tra due livelli di quantizzazione. In tal caso vi sono diverse strategie possibili, come arrotondare sempre in su, o sempre in giù, o a caso.

³ Nella (9.5) assumiamo che un numero situato a mezza strada fra due livelli di quantizzazione viene arrotondato in su.

mobile, l'errore relativo è più importante dell'errore assoluto. Ciò significa che gli errori in virgola mobile sono moltiplicativi anziché additivi. In altre parole, per la virgola mobile, se x rappresenta il valore prima del troncamento o arrotondamento e $Q(x)$ il valore dopo, si ha

$$Q[x] = x(1 + \varepsilon)$$

dove ε è l'errore relativo.

Per il caso dell'arrotondamento, per esempio, l'errore della mantissa è fra $\pm 0.5 \cdot 2^{-b}$, e di conseguenza l'errore nel valore della parola in virgola mobile è

$$-2^c \cdot 2^{-b}/2 < Q(x) - x \leq 2^c \cdot 2^{-b}/2$$

oppure, poiché $[Q(x) - x] = \varepsilon x$,

$$-2^c \cdot \frac{2^{-b}}{2} < \varepsilon x \leq 2^c \cdot \frac{2^{-b}}{2} \quad (9.7)$$

Pertanto, essendo $2^{c-1} \leq x < 2^c$, possiamo scrivere per il caso dell'arrotondamento

$$-2^{-b} < \varepsilon \leq 2^{-b} \quad (9.8a)$$

In modo analogo (v. il probl. 3 di questo capitolo) si può dimostrare che, per i casi complemento a uno e modulo e segno, il troncamento della mantissa dà luogo a

$$-2 \cdot 2^{-b} < \varepsilon \leq 0 \quad (9.8b)$$

mentre lo stesso troncamento della mantissa per il caso complemento a due dà luogo a

$$\begin{array}{ll} -2 \cdot 2^{-b} < \varepsilon \leq 0, & x > 0 \\ 0 \leq \varepsilon < 2 \cdot 2^{-b}, & x < 0 \end{array} \quad (9.8c)$$

9.2 LA QUANTIZZAZIONE NEL CAMPIONAMENTO DI SEGNALI ANALOGICI

I risultati del par. 9.1 possono applicarsi nell'analisi degli effetti della quantizzazione quando si campiona un segnale analogico. Nel par. 1.7, a proposito del campionamento, si fece l'ipotesi che fosse possibile ottenere una sequenza

$$x(n) = x_a(nT), \quad -\infty < n < \infty$$

dove $x_a(t)$ indica un segnale analogico a banda limitata. Si assunse implicitamente, cioè, che fosse possibile conoscere i campioni di $x_a(t)$ con precisione infinita. In tal caso, per rappresentare ogni campione sarebbe teoricamente necessario un numero infinito di bit. Ovviamente limitazioni fisiche impediscono di campionare con precisione infinita, e perciò ogni campione deve essere o troncato o arrotondato per poter stare in un registro di lunghezza finita. Pertanto una rappresentazione in qualche modo meno idealizzata del processo di campionamento è presentata nella fig. 9.2(a).

La forma della caratteristica del quantizzatore dipende da come sono rappresentati i campioni negativi e dal fatto che si usi l'arrotondamento o il troncamento. In particolare, assumiamo che i campioni di uscita siano rappresentati come frazioni a virgola fissa in complemento a due e lunghezza $(b + 1)$ bit. Si assuma inoltre che gli esatti campioni di ingresso $x(n)$ siano arrotondati al più vicino livello di quantizzazione, dando così luogo ai campioni quantizzati $\hat{x}(n)$. Per essere sicuri che i campioni non quantizzati siano contenuti nel campo di valori corrispondente ai $(b + 1)$ bit, è necessario anche assumere che la forma d'onda analogica sia normalizzata, così che

$$\left(-1 + \frac{2^{-b}}{2}\right) < x_a(nT) < \left(1 - \frac{2^{-b}}{2}\right) \quad (9.9)$$

(a)

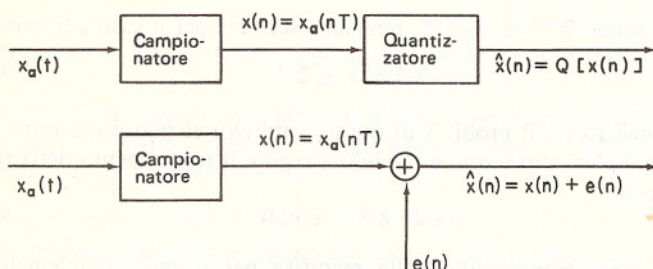


Fig. 9.2 Rappresentazione del campionamento di un segnale analogico: (a) modello non lineare; (b) modello statistico

Sotto queste condizioni la funzione del quantizzatore è come quella mostrata nella fig. 9.3, dove si è assunto $b = 2$.

Se il valore esatto di un campione di ingresso cade al di fuori del campo indicato nella (9.9) si ha un'ulteriore distorsione. Come mostrato nella fig. 9.3, a tutti i campioni che superano il valore $1 - 0.5 \cdot 2^{-b}$ si assegna il valore quantizzato $1 - 2^{-b}$, e a tutti i campioni minori di $-(1 + 0.5 \cdot 2^{-b})$ si assegna il valore -1 . Ovviamente questo taglio dell'ingresso non è in generale desiderabile e deve essere eliminato riducendo l'ampiezza dell'ingresso finché la (9.9) non sia soddisfatta.

Una rappresentazione equivalente del processo di quantizzazione è mostrata nella fig. 9.2(b). Essa indica che i valori quantizzati si possono esprimere come

$$\hat{x}(n) = Q[x(n)] = x(n) + e(n)$$

dove $x(n)$ è il campione esatto ed $e(n)$ è chiamato l'errore di quantizzazione. Poiché si è supposto di arrotondare, è

$$-\frac{\Delta}{2} < e(n) \leq \frac{\Delta}{2}$$

dove Δ è l'intervallo di quantizzazione, ovvero $\Delta = 2^{-b}$.

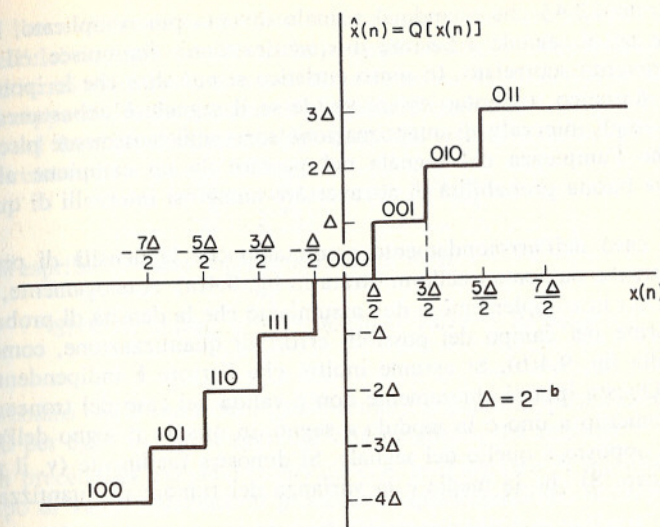


Fig. 9.3 Arrotondamento in complemento a due di campioni per $b = 2$

Affinché la fig. 9.2(b) sia perfettamente equivalente alla fig. 9.2(a), occorre ovviamente conoscere in modo esatto $e(n)$ per ogni n . Nella maggior parte dei casi è ragionevole assumere che $e(n)$ non è noto e allora, per rappresentare gli effetti della quantizzazione nel campionamento, può essere utile un modello statistico basato sulla fig. 9.2(b). Noi useremo un tale modello anche per descrivere gli effetti della quantizzazione negli algoritmi di elaborazione dei segnali. In particolare si fanno comunemente le seguenti ipotesi:

1. La sequenza dei campioni di errore $\{e(n)\}$ è una sequenza campione di un processo casuale stazionario.
2. La sequenza errore è scorrelata con la sequenza dei campioni esatti $\{x(n)\}$.
3. Le variabili casuali del processo errore sono scorrelate e cioè l'errore è un rumore bianco.
4. La densità di probabilità del processo errore è uniforme nel campo di valori dell'errore di quantizzazione.

Queste ipotesi, che sono soprattutto degli espedienti per poter trattare matematicamente il problema, portano, come vedremo, ad un'analisi alquanto semplice degli effetti della quantizzazione. È facile trovare esempi dove queste ipotesi chiaramente non valgono. Per esempio, se $x_a(t)$ è una funzione a gradino, risulta impossibile giustificare le ipotesi fatte sopra. Tuttavia queste stesse ipotesi diventano più realistiche quando la sequenza $x(n)$ è un segnale complicato, come il parlato o la musica, dove il segnale fluttua rapidamente in modo alquanto imprevedibile. Gli esperimenti han-

no mostrato [2-4] che quando il segnale diventa più complicato, la correlazione tra il segnale e l'errore di quantizzazione diminuisce, ed anche l'errore diventa scorrelato. In senso euristico si può dire che le ipotesi del modello statistico sembrano essere valide se il segnale è abbastanza complesso e se gli intervalli di quantizzazione sono sufficientemente piccoli, in modo che l'ampiezza del segnale nel passare da un campione all'altro abbia una buona probabilità di attraversare numerosi intervalli di quantizzazione.

Nel caso dell'arrotondamento assumiamo che la densità di probabilità dell'errore sia come quella mostrata in fig. 9.4(a). Analogamente, per il troncamento in complemento a due assumiamo che la densità di probabilità sia uniforme nel campo dei possibili errori di quantizzazione, come mostrato nella fig. 9.4(b). Si assume inoltre che l'errore è indipendente dal segnale. Questa ipotesi chiaramente non è valida nel caso del troncamento in complemento a uno o in modulo e segno, in quanto il segno dell'errore è sempre opposto a quello del segnale. Si dimostra facilmente (v. il probl. n. 3 del cap. 8) che la media e la varianza del rumore di quantizzazione valgono

$$m_e = 0$$

$$\sigma_e^2 = \frac{\Delta^2}{12} = \frac{2^{-2b}}{12}$$

per l'arrotondamento, e

$$m_e = -\frac{2^{-b}}{2}$$

$$\sigma_e^2 = \frac{2^{-2b}}{12}$$

per il troncamento in complemento a due. La sequenza di autocovarianza dell'errore si assume uguale a $\gamma_{ee}(n) = \sigma_e^2 \delta(n)$

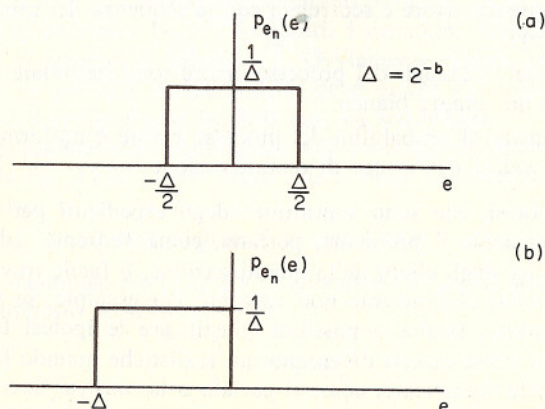


Fig. 9.4 Funzioni di densità di probabilità per (a) arrotondamento; (b) troncamento

sia per l'arrotondamento che per il troncamento in complemento a due.

Nell'elaborazione numerica dei segnali analogici campionati, l'errore di quantizzazione è comunemente visto come un segnale di rumore additivo. Il rapporto fra la potenza del segnale e quella del rumore è un'utile misura del loro peso relativo. Nel caso dell'arrotondamento il rapporto segnale-rumore è

$$\frac{\sigma_x^2}{\sigma_e^2} = \frac{\sigma_x^2}{2^{-2b}/12} = (12 \cdot 2^{2b}) \sigma_x^2$$

Se lo si esprime in scala logaritmica, come nella

$$\text{SNR} = 10 \log_{10} \left(\frac{\sigma_x^2}{\sigma_e^2} \right) = 6.02b + 10.79 + 10 \log_{10} (\sigma_x^2) \quad (9.10)$$

risulta chiaro che il rapporto segnale-rumore aumenta approssimativamente di 6 dB per ogni bit che si aggiunge alla lunghezza del registro.

In precedenza abbiamo osservato che, se il segnale di ingresso eccede il campo di valori associato al processo di quantizzazione, occorre ridurre l'ampiezza dell'ingresso per evitare che il segnale venga « tagliato ». In altri termini, si campiona $Ax(n)$ invece di $x(n)$, dove $0 < A < 1$. Poiché la varianza di $Ax(n)$ è $A^2 \sigma_x^2$, il rapporto segnale-rumore diventa

$$\text{SNR} = 10 \log_{10} \left(\frac{A^2 \sigma_x^2}{\sigma_e^2} \right) = 6b + 10.8 + 10 \log_{10} (\sigma_x^2) + 20 \log_{10} (A) \quad (9.11)$$

Un confronto tra la (9.10) e la (9.11) mostra che la riduzione dell'ampiezza dell'ingresso, fatta allo scopo di evitare la distorsione da taglio, riduce il rapporto segnale-rumore. Molti segnali analogici, quali il parlato o la musica, possono essere convenientemente visti come processi casuali. Generalmente tali segnali sono caratterizzati da densità di probabilità che presentano un picco in prossimità dello zero e diminuiscono poi rapidamente al crescere dell'ampiezza. In tali casi la probabilità che l'ampiezza di un dato campione superi di tre o quattro volte la radice del valore quadratico medio del segnale è molto bassa. Pertanto, se si pone A pari a $1/(4 \sigma_x)$, allora con elevata probabilità non si verificherà distorsione da taglio. In tal caso il rapporto segnale-rumore diventa

$$\text{SNR} = 6b - 1.24 \text{ dB}$$

Ne segue che, per avere $\text{SNR} \geq 80 \text{ dB}$, deve essere $b = 14$ bit. Questa relazione fra la dinamica del segnale e l'errore di quantizzazione è una caratteristica fondamentale dell'uso della virgola fissa nella elaborazione dei segnali a tempo discreto. Vedremo infatti ripresentarsi questa relazione lungo tutto il resto del presente capitolo.

Quando si elaborano segnali quantizzati, l'errore (o rumore) di ingresso si manifesta come un errore (o rumore) nell'uscita risultante. Per esempio, se una sequenza quantizzata $\hat{x}(n) = x(n) + e(n)$ è l'ingresso a un sistema lineare invariante alla traslazione, allora l'uscita può essere rappresentata come $\hat{y}(n) = y(n) + f(n)$, dove $y(n)$ è la risposta ad $x(n)$ e

$f(n)$ è la risposta ad $e(n)$. Poiché $x(n)$ ed $e(n)$ sono indipendenti, $x(n)$ può essere ignorato nel calcolare la potenza del rumore di uscita. Usando le (8.50) e (8.53) e il fatto che il rumore di ingresso è bianco, possiamo esprimere la media e la varianza del rumore di uscita come

$$m_f = m_e \sum_{n=-\infty}^{\infty} h(n) = m_e H(e^{j0}) \quad (9.12)$$

$$\sigma_f^2 = \sigma_e^2 \sum_{n=-\infty}^{\infty} |h(n)|^2 = \frac{\sigma_e^2}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega \quad (9.13)$$

Nel derivare queste espressioni, abbiamo implicitamente assunto che il sistema è stato realizzato senza errori. In effetti questo non è vero; è tuttavia ragionevole ammettere che gli errori in uscita dovuti ad errori introdotti nella realizzazione del sistema sono indipendenti dagli errori dovuti al rumore di quantizzazione dell'ingresso. Pertanto, gli errori dovuti all'arrotondamento o al troncamento nella realizzazione di un filtro numerico sono considerati a parte, e i loro effetti vengono quindi sovrapposti agli errori dovuti alla quantizzazione dell'ingresso.

9.3 EFFETTI DELLA LUNGHEZZA FINITA DEI REGISTRI NELLA REALIZZAZIONE DI FILTRI NUMERICI IIR

In questo paragrafo discuteremo l'effetto della quantizzazione dei risultati di operazioni aritmetiche nella realizzazione dei filtri numerici IIR. Come già osservato nel cap. 4, le operazioni aritmetiche di base necessarie alla realizzazione di un filtro numerico sono la moltiplicazione per una costante (i coefficienti del filtro) e l'addizione. Per l'aritmetica in virgola fissa, il risultato di una moltiplicazione deve essere arrotondato o troncato, ma non è così per il risultato di un'addizione. Tuttavia, poiché il risultato di una somma può avere una lunghezza maggiore di quella del registro, ne discendono importanti conseguenze sulla dinamica dei segnali. Il problema è simile a quelli connessi alla quantizzazione dei campioni di un segnale analogico, dove abbiamo già visto che le richieste di un'ampia dinamica di segnale e di un piccolo errore di quantizzazione sono in contrasto fra loro. Non è così per le realizzazioni in virgola mobile, dove si hanno limitazioni di dinamica molto meno severe, ma occorre introdurre il troncamento o l'arrotondamento sia per la moltiplicazione che per l'addizione. Come si è visto nel precedente paragrafo, il troncamento o l'arrotondamento sono processi non lineari. In altri termini, l'effetto della quantizzazione nella realizzazione di un filtro numerico invariante alla traslazione è quello di introdurre elementi non lineari in alcuni rami della struttura del filtro.

In definitiva, le trasformazioni realizzate in pratica con i sistemi lineari invarianti alla traslazione sono generalmente non lineari e quindi affette da

errori. Per potere valutare questi errori ed anche per poterli contenere entro limiti accettabili (cercando al tempo stesso di minimizzare la complessità dell'« hardware » necessario alla realizzazione dei filtri), è importante comprendere gli effetti non lineari originati dalla quantizzazione. L'analisi di questi effetti resta per molti casi ancora da chiarire, ed è per lo più complicata al punto da richiedere dettagli che non possono trovar posto in questo libro.

Nella discussione che segue considereremo innanzitutto alcuni semplici casi in aritmetica a virgola fissa dove la non linearità causa un errore periodico all'uscita se l'ingresso è nullo, o pari a una costante, o sinusoidale. In tali semplici casi è possibile comprendere fin nei dettagli gli effetti della non linearità. Tuttavia, quando l'ingresso non è né costante, né sinusoidale, risulta conveniente usare il modello statistico del paragrafo precedente sostituendo al sistema non lineare un sistema lineare con sorgenti interne di rumore additivo. Usando questo modello è infatti possibile, per ogni filtro dato, calcolare un errore medio in uscita dovuto alla quantizzazione dei risultati delle operazioni aritmetiche implicite nella realizzazione del filtro stesso.

9.3.1 Cicli limite per ingresso zero nelle realizzazioni in virgola fissa di filtri numerici IIR

Se un filtro numerico stabile realizzato con un'aritmetica a precisione infinita ha una eccitazione nulla per n maggiore di un certo valore n_0 , l'uscita del filtro tenderà asintoticamente a zero per $n > n_0$. Per lo stesso filtro, realizzato con un'aritmetica a lunghezza finita di registro, l'uscita potrà smorzarsi fino a un campo di valori di ampiezza non nulla ed avere poi un comportamento oscillatorio. Questo effetto è chiamato spesso *comportamento di ciclo limite per ingresso zero* ed è una conseguenza dei quantizzatori non lineari nell'anello di reazione del filtro. Il comportamento di ciclo limite di un filtro numerico è complesso e difficile da analizzare, e noi non tenteremo nemmeno di trattare l'argomento in maniera generale.

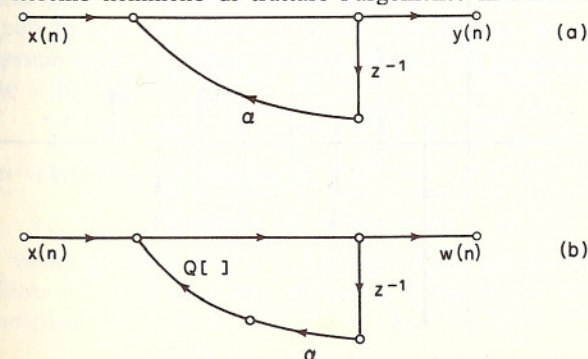


Fig. 9.5 Grafi di flusso per un sistema IIR del primo ordine: (a) sistema lineare ideale; (b) sistema non lineare dovuto alla quantizzazione del prodotto

Per filtri semplici del primo e secondo ordine è possibile tuttavia capire l'effetto e dare una interpretazione delle oscillazioni in termini dello spostamento dei poli effettivi del filtro sulla circonferenza unitaria [5-10]. L'effetto verrà meglio illustrato per mezzo di un esempio.

ESEMPIO. Per un'illustrazione degli effetti di ciclo limite si consideri il sistema del primo ordine caratterizzato dall'equazione alle differenze

$$y(n) = \alpha y(n-1) + x(n) \quad (9.14)$$

Il grafo di flusso di segnale di questo sistema è mostrato nella fig. 9.5(a). Assumiamo che sia $\alpha = 0.5$ e che la lunghezza di registro per memorizzare il coefficiente α , l'ingresso $x(n)$, e la variabile di nodo del filtro $y(n-1)$, sia di quattro bit (vale a dire un bit segno alla sinistra della virgola binaria a tre bit alla destra). A causa dei registri di lunghezza finita, il prodotto $\alpha y(n-1)$ deve essere arrotondato o troncato a

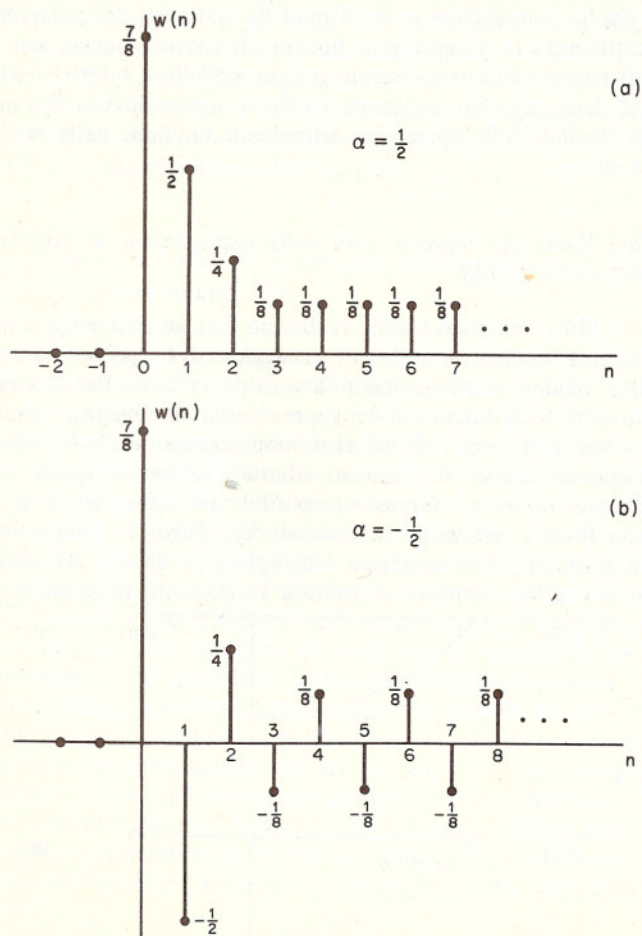


Fig. 9.6 Risposta di un sistema quantizzato del primo ordine a un campione unitario (a) $\alpha = 0.5$; (b) $\alpha = -0.5$

quattro bit prima di essere sommato a $x(n)$. Il grafo di flusso che rappresenta l'effettiva realizzazione basata sulla (9.14) è mostrato nella fig. 9.5(b). Se si suppone che il prodotto venga arrotondato, l'uscita effettiva $w(n)$ soddisfa l'equazione alle differenze non lineare

$$w(n) = Q[\alpha w(n-1)] + x(n) \quad (9.15)$$

dove $Q[\]$ rappresenta l'operazione di arrotondamento. Assumiamo che sia $\alpha = 0.5 = 0.100$ e che l'ingresso sia un campione unitario di ampiezza $7/8 = 0.111$. Usando la (9.15) vediamo che per $n = 0$, $w(0) = 7/8 = 0.111$. Per ottenere $w(1)$ moltiplichiamo $w(0)$ per α , ottenendo il risultato $\alpha w(0) = 0.011100$, un numero a sette bit che deve essere arrotondato a quattro bit. Questo numero, $7/16$, è esattamente a metà fra i due livelli di quantizzazione a quattro bit $4/8$ e $3/8$. Se in tali casi scegliamo di arrotondare sempre in su, allora 0.011100 arrotondato a quattro bit è $0.100 = 1/2$. Poiché $x(1) = 0$, sarà $w(1) = 0.100 = 1/2$. Continuando si ha $w(2) = Q[\alpha w(1)] = 0.010 = 1/4$ e $w(3) = 0.001 = 1/8$. In entrambi questi casi non è necessario alcun arrotondamento. Tuttavia, per ottenere $w(4)$, dobbiamo arrotondare il numero a sette bit $\alpha w(3) = 0.000100$ a 0.001 . Lo stesso risultato si ottiene per tutti i valori di $n \geq 3$. La sequenza di uscita per questo esempio è mostrata nella fig. 9.6(a). Se $\alpha = -0.5$ i calcoli svolti sopra possono essere ripetuti mostrando che l'uscita è come quella della fig. 9.6(b). Pertanto, a causa dell'arrotondamento del prodotto $\alpha w(n-1)$, l'uscita raggiunge il valore costante di $1/8$ quando $\alpha = 0.5$ e un'oscillazione periodica di regime fra $+1/8$ e $-1/8$ quando $\alpha = -0.5$. Tali uscite periodiche sono simili a quelle che si otterrebbero da un filtro del primo ordine con un polo in $z = \pm 1$ invece che in $\pm \alpha$. Quando $\alpha = +0.5$ il periodo di oscillazione è 1, e quando $\alpha = -0.5$ il periodo di oscillazione è 2. Queste uscite periodiche a regime sono chiamate *cicli limite*, e la loro esistenza fu notata per primo da Blackman [11], il quale denominò *bande morte* gli intervalli di ampiezza cui tali cicli limite sono confinati. Nel nostro caso la banda morta è $-2^{-b} \leq w(n) \leq 2^{-b}$.

La possibile esistenza di un ciclo limite per ingresso zero è importante per quelle applicazioni dove un filtro numerico deve essere attivo continuamente, in quanto generalmente si desidera che l'uscita approssimi zero se l'ingresso è zero. Per esempio, si consideri un segnale voce campionato che viene filtrato con un filtro numerico e poi riportato ad essere un segnale acustico mediante un convertitore numerico-analogico. In una tale situazione sarebbe assai poco desiderabile che il filtro entrasse in un ciclo limite periodico tutte le volte che l'ingresso è zero.

Jackson [5,6] ha considerato il comportamento di ciclo limite nei sistemi del primo e secondo ordine, mediante un'analisi basata sull'osservazione fatta sopra e, cioè, che nel ciclo limite il sistema si comporta come se i suoi poli fossero sul circolo unitario. Considerando in particolare il filtro del primo ordine di sopra, osserviamo che per definizione di arrotondamento si ha

$$|Q[\alpha w(n-1)] - \alpha w(n-1)| \leq \frac{1}{2} \cdot 2^{-b} \quad (9.16)$$

Inoltre, per valori di n nel ciclo limite,

$$|Q[\alpha w(n-1)]| = |w(n-1)|$$

e, cioè, il valore effettivo di α è 1, corrispondente allo spostamento del polo del filtro sul circolo unitario. Il campo di valori per cui si realizza questa condizione è

$$|w(n-1)| - |\alpha w(n-1)| \leq \frac{1}{2} \cdot 2^{-b}$$

oppure, risolvendo rispetto a $|w(n-1)|$,

$$|w(n-1)| \leq \frac{\frac{1}{2} \cdot 2^{-b}}{1 - |\alpha|} \quad (9.17)$$

La (9.17) definisce la banda morta per il filtro del primo ordine. Come conseguenza dell'arrotondamento, i valori interni alla banda morta sono quantizzati in intervalli di 2^{-b} . Si noti che la (9.17) dà il valore corretto per la banda morta quando $|\alpha| = 0.5$. Ogni qualvolta la variabile di nodo $w(n-1)$ cade internamente alla banda morta quando l'ingresso è zero, il filtro entra in un ciclo limite e vi rimane finché non viene applicato un ingresso che riporta l'uscita fuori della banda morta.

Per un filtro del secondo ordine esiste una grande varietà di comportamenti di ciclo limite. Si consideri l'equazione alle differenze del secondo ordine

$$y(n) = x(n) + \alpha_1 y(n-1) + \alpha_2 y(n-2) \quad (9.18)$$

Con $\alpha_1^2 < -4\alpha_2$, i poli del filtro sono complessi coniugati, e, con $\alpha_2 = -1$, i poli sono sul circolo unitario. Tenendo conto della lunghezza finita dei registri la (9.18) si può riscrivere

$$w(n) = x(n) + Q[\alpha_1 w(n-1)] + Q[\alpha_2 w(n-2)] \quad (9.19)$$

dove $Q[\]$ rappresenta ancora l'arrotondamento dei prodotti indicati. Come prima, per definizione di arrotondamento, si ha

$$|Q[\alpha_2 w(n-2)] - \alpha_2 w(n-2)| \leq \frac{1}{2} \cdot 2^{-b} \quad (9.20)$$

Con $x(n) = 0$, i poli del sistema si sposteranno sul circolo unitario se

$$Q[\alpha_2 w(n-2)] = w(n-2)$$

Sostituendo questa espressione nella (9.20) otteniamo

$$|w(n-2)| - |\alpha_2 w(n-2)| \leq \frac{1}{2} \cdot 2^{-b}$$

oppure, risolvendo rispetto a $|w(n-2)|$,

$$|w(n-2)| \leq \frac{\frac{1}{2} \cdot 2^{-b}}{1 - |\alpha_2|} \quad (9.21)$$

Pertanto, se $w(n-2)$ cade in questo campo di valori quando l'ingresso è zero, il valore effettivo di α_2 è tale per cui i poli del sistema sono proprio sul circolo unitario. Sotto queste condizioni, il valore di α_1 controlla la frequenza dell'oscillazione.

In una seconda modalità di comportamento di ciclo limite che può verificarsi in filtri del secondo ordine, l'effetto dell'arrotondamento è quello di spostare i poli in $z = +1$ e $z = -1$. La banda morta corrispondente a questa modalità è limitata da $1/(1 - |\alpha_1| - \alpha_2)$. L'ampiezza di un ciclo limite in questa banda è ovviamente quantizzata in intervalli di 2^{-b} [5, 6].

Oltre alle classi di cicli limite viste sopra, un tipo più grave di ciclo limite può verificarsi a causa della saturazione. L'effetto della saturazione è quello di inserire un notevole errore in uscita, e, in alcuni casi, da quel

punto in poi l'uscita del filtro oscilla tra i limiti di massima ampiezza. Tali cicli limite sono stati chiamati *oscillazioni da saturazione*. Il problema delle oscillazioni causate da saturazione è discusso in dettaglio da Ebert ed altri [12]. Un esempio semplice è sviluppato nel probl. 9 di questo capitolo.

La discussione svolta fin qui ha riguardato soltanto i cicli limite per ingresso zero, causati dall'arrotondamento nei sistemi IIR del primo e secondo ordine. Benché l'analisi sia stata alquanto euristica, l'esperienza ha dimostrato che le semplici formule così ricavate sono coerenti con i risultati sperimentali e sono utili per predire i comportamenti di ciclo limite nei filtri numerici IIR. Un tipo simile di analisi può essere svolto per il caso del troncamento (si veda, per esempio, il probl. 6 alla fine di questo capitolo). Nel caso dei sistemi di ordine più elevato realizzati con struttura in parallelo, le uscite dei singoli sistemi del secondo ordine sono indipendenti quando l'ingresso è zero. Pertanto si può applicare direttamente l'analisi svolta sopra. Nel caso di realizzazioni in cascata soltanto la prima sezione ha ingresso zero. Le sezioni successive possono mostrare il loro caratteristico comportamento di ciclo limite oppure possono vedersi come semplicemente filtranti l'uscita di ciclo limite di una sezione precedente. Per sistemi di ordine elevato realizzati con altre strutture, il comportamento di ciclo limite diventa più complesso e con esso la sua analisi. Quando l'ingresso è diverso da zero, gli effetti della quantizzazione dipendono dall'ingresso e il tipo di analisi adottato in questo paragrafo diventa del tutto inadeguato eccetto che per ingressi semplici quali un campione unitario, un gradino unitario o una sinusoide. Negli altri casi la complessità dei fenomeni di quantizzazione ci costringe ad adottare un modello statistico.

I risultati ricavati sopra, oltre a spiegare gli effetti di ciclo limite nei filtri numerici, risultano utili quando l'uscita desiderata di un sistema è proprio la risposta di ciclo limite per ingresso zero. È questo il caso, per esempio, quando si è interessati agli oscillatori sinusoidali numerici per la generazione di segnali e per la generazione dei coefficienti nelle trasformazioni di Fourier discrete.

9.3.2 Analisi statistica della quantizzazione nelle realizzazioni in virgola fissa di filtri numerici IIR

Un'analisi precisa degli errori di troncamento o arrotondamento non è generalmente richiesta nelle applicazioni pratiche. Per esempio, un obiettivo usuale nell'analisi degli errori è quello di scegliere la lunghezza di registro necessaria a soddisfare alcune specifiche sul peso relativo fra segnale ed errori. La lunghezza di registro può, ovviamente, essere cambiata soltanto a intervalli di un bit. Come vedremo, l'aggiunta di un bit alla lunghezza di registro riduce l'ampiezza degli errori di quantizzazione approssimativamente di un fattore un mezzo. Pertanto una decisione finale circa la lunghezza di registro è poco sensibile ad imprecisioni nell'analisi

degli errori; spesso una analisi corretta al 30-40% risulta adeguata. Proprio in ragione di questa scarsa sensibilità è possibile utilizzare per l'analisi degli errori di quantizzazione il modello statistico sviluppato nel par. 9.2.

Nella discussione che segue tratteremo soprattutto semplici sistemi del primo e del secondo ordine per illustrare il modo in cui può essere usato il modello statistico per stimare gli effetti della quantizzazione nella realizzazione di filtri numerici in virgola fissa. Da questi semplici esempi emergeranno diversi criteri che potranno generalizzarsi ad esempi più complessi. Tali criteri torneranno utili al fine di valutare i molti pro e contro nella ricerca delle più efficienti ed economiche realizzazioni di filtri numerici.

Si consideri una realizzazione in virgola fissa di un sistema del primo ordine con arrotondamento applicato ai prodotti. La fig. 9.7(a) rappresenta il sistema con precisione infinita e la fig. 9.7(b) la realizzazione con precisione finita, con $Q[\]$ che indica l'operazione di arrotondamento.

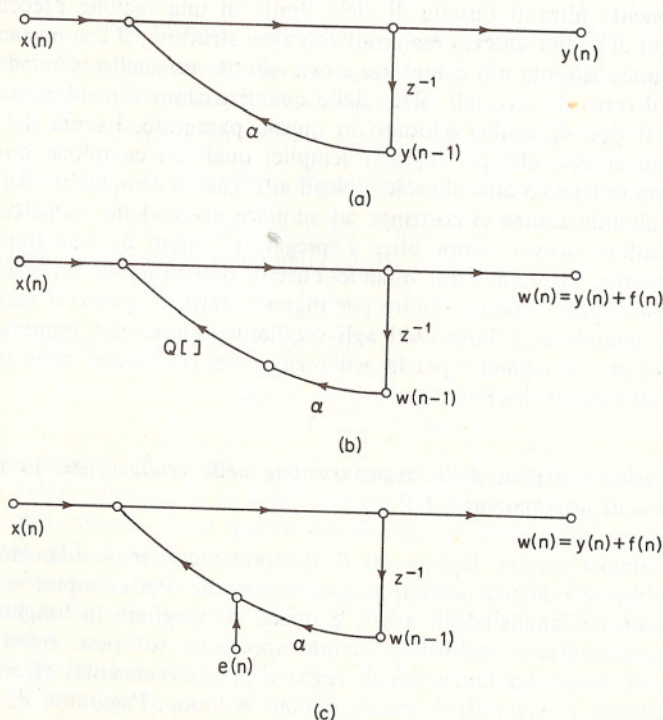


Fig. 9.7 Grafi di flusso per sistemi IIR del primo ordine: (a) sistema lineare ideale; (b) sistema non lineare; (c) modello statistico per rumore di arrotondamento in virgola fissa

Nella fig. 9.7(c) è illustrato lo stesso sistema, ma con l'effetto del quantizzatore rappresentato mediante la sorgente di rumore additivo

$$e(n) = Q[\alpha w(n-1)] - \alpha w(n-1)$$

Come già osservato nel par. 9.2, le rappresentazioni di fig. 9.7(b) e (c) sono identiche quando $e(n)$ è noto. Faremo invece le seguenti ipotesi circa l'effetto della quantizzazione dei prodotti.

1. La sequenza errore $e(n)$ è una sequenza rumore bianco.
2. La sequenza errore ha una densità di probabilità uniforme su un intervallo di quantizzazione.
3. La sequenza errore $e(n)$ è scorrelata con l'ingresso $x(n)$ e con $\alpha w(n-1)$. Ciò implica che $e(n)$ è scorrelata con l'uscita (si veda il probl. 12 del cap. 8).

Queste ipotesi sono identiche a quelle fatte per la quantizzazione dei campioni di un segnale analogico, e le condizioni per la loro validità pressoché le stesse. Vale a dire che queste ipotesi valgono quando il segnale di ingresso e le risultanti variabili di nodo variano da campione a campione in maniera sufficientemente complessa. Esse chiaramente non valgono per ingressi quali un campione unitario, un gradino unitario o una sequenza sinusoidale.

Se la lunghezza di registro è $(b+1)$ bit, allora, per l'arrotondamento

$$-\frac{1}{2} \cdot 2^{-b} < e(n) \leq \frac{1}{2} \cdot 2^{-b}$$

Assumendo una distribuzione uniforme su questo campo di valori, la media di $e(n)$ è zero e

$$\sigma_e^2 = \frac{1}{12} \cdot 2^{-2b}$$

Se $y(n)$ è l'uscita che si otterrebbe da $x(n)$ senza errori di quantizzazione, l'uscita effettiva può essere rappresentata come

$$w(n) = y(n) + f(n)$$

dove $f(n)$ rappresenta l'errore di uscita dovuto alla sorgente di rumore $e(n)$. Se $h_e(n)$ è la risposta all'impulso del sistema compreso fra il nodo nel quale entra $e(n)$ e l'uscita, allora

$$m_f = m_e \sum_{n=-\infty}^{\infty} h_e(n) \quad (9.22a)$$

e poiché $e(n)$ è per ipotesi rumore bianco⁴,

$$\sigma_f^2 = \sigma_e^2 \sum_{n=-\infty}^{\infty} h_e^2(n) \quad (9.22b)$$

Dalla fig. 9.7(c) vediamo che per questo caso la risposta al campione unitario dall'ingresso del rumore all'uscita è la stessa che per il segnale. Ciò

⁴ Facciamo l'ipotesi per convenienza che $h_e(n)$ sia reale.

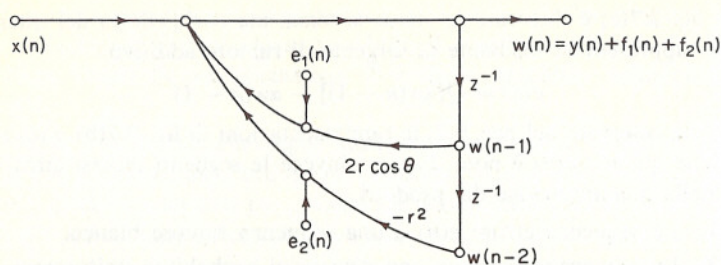


Fig. 9.8 Modello statistico per rumore di arrotondamento in virgola fissa in un sistema IIR del secondo ordine.

ovviamente non è vero in generale. Per questo esempio, dunque, $h_e(n) = \alpha^n u(n)$, così che

$$\sigma_f^2 = \sigma_e^2 \frac{1}{1 - \alpha^2} = \frac{1}{12} \cdot 2^{-2b} \frac{1}{1 - \alpha^2} \quad (9.23)$$

Come secondo esempio consideriamo un filtro del secondo ordine con una coppia di poli complessi in $z = re^{\pm j\theta}$, corrispondente all'equazione alle differenze

$$y(n) = x(n) + 2r \cos \theta y(n-1) - r^2 y(n-2)$$

Con l'arrotondamento dei prodotti si ottiene l'equazione alle differenze non lineare

$$w(n) = x(n) + Q[2r \cos \theta w(n-1)] - Q[r^2 w(n-2)]$$

Essendoci due moltiplicazioni, si introducono due sorgenti di rumore come indicato nella fig. 9.8. Queste sorgenti sono indicate con $e_1(n)$ ed $e_2(n)$. Analogamente a prima, l'uscita può essere rappresentata come la somma dell'uscita ideale e di due componenti di errore, $f_1(n)$ ed $f_2(n)$, dovute rispettivamente ad $e_1(n)$ ed $e_2(n)$. Come prima, assumiamo che $e_1(n)$ ed $e_2(n)$ siano sequenze rumore bianco, con densità di ampiezza uniformi tra $\pm 0.5 \cdot 2^{-b}$ ed entrambe scorrelate con l'ingresso. Faremo inoltre l'ipotesi che siano scorrelate tra loro. Da ciò segue che anche $f_1(n)$ ed $f_2(n)$ sono scorrelate (v. il probl. 12 del cap. 8) e di conseguenza (v. probl. 2 del cap. 8)

$$\sigma_f^2 = \sigma_{f_1}^2 + \sigma_{f_2}^2$$

oppure, chiamando con $h_1(n)$ ed $h_2(n)$ le risposte al campione unitario dagli ingressi delle sorgenti di rumore.

$$\sigma_f^2 = \sigma_{e_1}^2 \sum_{n=-\infty}^{\infty} h_1^2(n) + \sigma_{e_2}^2 \sum_{n=-\infty}^{\infty} h_2^2(n)$$

Osserviamo che, per questo esempio, $h_1(n)$ ed $h_2(n)$ sono uguali e sono date da

$$h_1(n) = h_2(n) = \frac{1}{\sin \theta} r^n \sin [(n+1)\theta] u(n)$$

Si può verificare che

$$\sum_{n=-\infty}^{\infty} h_1^2(n) = \frac{1+r^2}{1-r^2 r^4 + 1 - 2r^2 \cos 2\theta}$$

così che

$$\begin{aligned} \sigma_{e_1}^2 &= \sigma_{e_2}^2 = \frac{1}{12} \cdot 2^{-2b} \\ \sigma_f^2 &= \frac{2}{12} \cdot 2^{-2b} \frac{1+r^2}{1-r^2 r^4 + 1 - 2r^2 \cos 2\theta} \end{aligned} \quad (9.24)$$

Questi due esempi illustrano un tipo di analisi che può applicarsi ad ogni filtro in virgola fissa per ottenere la varianza del rumore di uscita dovuto al processo di arrotondamento. Alcuni altri esempi sono presi in considerazione nei problemi alla fine del capitolo.

I risultati di sopra si modificano facilmente nel caso del troncamento in complemento a due. Osserviamo che l'ampiezza dell'errore nel caso del troncamento in complemento a due cade nel campo di valori

$$-2^{-b} < E_T \leq 0$$

Pertanto, se si vuole rappresentare l'effetto del troncamento in complemento a due in modo analogo a quanto fatto per l'arrotondamento, occorre considerare una sorgente di rumore additivo $e(n)$ con una densità di probabilità di ampiezza uniforme tra -2^{-b} e zero. Si assume ancora che $e(n)$ è linearmente indipendente da se stesso traslato (rumore bianco) ma esso non ha più media nulla. In particolare, $E[e(n)] = -0.5 \cdot 2^{-b}$. Tuttavia, la varianza di $e(n)$ in questo caso è identica a quella del caso dell'arrotondamento. Perciò, con il troncamento in complemento a due, la varianza del rumore di uscita è la stessa che per l'arrotondamento. Il valor medio del rumore di uscita, tuttavia, non è più nullo, e lo si può calcolare facilmente usando la (9.22a). In molti esempi i risultati fin qui trovati possono considerarsi applicabili anche al caso del troncamento in complemento a uno e in modulo e segno, sebbene, come si è visto nel par. 9.2, in quei casi la correlazione fra il segnale e l'errore di troncamento sia, in effetti, maggiore, in quanto la polarità dell'errore è sempre opposta alla polarità del segnale al quale si applica il troncamento.

Come osservato in precedenza, un altro problema da tenere in considerazione nella realizzazione in virgola fissa dei filtri numerici è la possibilità della saturazione. Con la convenzione che ogni registro in virgola fissa rappresenta una frazione con segno, ogni nodo del filtro deve essere vincolato a mantenere un'ampiezza minore di uno onde evitare la saturazione. Indicando con $x(n)$ l'ingresso del filtro e con $y_k(n)$ e $h_k(n)$ rispettivamente l'uscita del k -mo. nodo e la risposta al campione unitario dall'ingresso del filtro fino al k -mo nodo, si ha

$$y_k(n) = \sum_{r=-\infty}^{\infty} h_k(r) x(n-r)$$

Se x_{\max} è il massimo del valore assoluto dell'ingresso, allora

$$|y_k(n)| \leq x_{\max} \sum_{r=-\infty}^{\infty} |h_k(r)| \quad (9.25)$$

Pertanto, poiché si vuole che sia $|y_k(n)| < 1$, per la (9.25) deve essere

$$x_{\max} < \frac{1}{\sum_{r=-\infty}^{\infty} |h_k(r)|} \quad (9.26)$$

per tutti i nodi della rete. La (9.26) fornisce quindi una limitazione superiore per il valore massimo dell'ingresso tale da assicurare che non si verifichi saturazione nel k -mo nodo. Nel caso più generale, per garantire che non si abbia saturazione, è necessario scalare l'ingresso proprio come indicato dalla (9.26). Ciò è conseguenza del fatto che nella (9.25) può anche valere il segno di uguaglianza come è il caso per la sequenza $x(n)$ tale che, per $n = n_0$, $x(n_0 - r) = \text{sgn}[h_k(r)]$ [dove $\text{sgn}(x) = 1$ per $x \geq 0$ e $\text{sgn}(x) = -1$ per $x < 0$]. La condizione (9.26) può essere soddisfatta applicando un'attenuazione al segnale all'ingresso del filtro.

Come esempio si consideri un ingresso $x(n)$ che è una sequenza rumore bianco con densità di probabilità di ampiezza uniforme. Per il caso del filtro del primo ordine occorre allora scegliere un'ampiezza massima di ingresso pari a $(1 - |\alpha|)$. Per questo caso, se σ_x^2 è la varianza del segnale di ingresso e σ_y^2 quella del segnale di uscita, si ha

$$\sigma_x^2 = \frac{1}{3} (1 - |\alpha|)^2 \quad (9.27)$$

$$\sigma_y^2 = \frac{1}{3} \frac{(1 - |\alpha|)^2}{1 - |\alpha|^2} \quad (9.28)$$

Per questo esempio possiamo allora calcolare un rapporto rumore-segnale di uscita mediante il rapporto σ_f^2/σ_y^2 , con il risultato che

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{4} \cdot 2^{-2b} \frac{1}{(1 - |\alpha|)^2} \quad (9.29)$$

In modo analogo possiamo ricavare un rapporto rumore-segnale per il filtro del secondo ordine considerato prima. Come per il caso del filtro del primo ordine, vincoleremo l'ampiezza dell'ingresso per garantire che non sia superata la dinamica dei registri. Se ammettiamo che la sequenza di ingresso sia rumore bianco uniformemente distribuito, il risultante rapporto rumore-segnale in uscita sarà

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{2} \cdot 2^{-2b} \left(\sum_{n=-\infty}^{\infty} |h(n)| \right)^2 = \frac{1}{2} \cdot 2^{-2b} \left(\frac{1}{\sin \theta} \sum_{n=0}^{\infty} r^n |\sin[(n+1)\theta]| \right)^2 \quad (9.30)$$

Sebbene sia difficile valutare esattamente questa espressione, è possibile ottenere una limitazione superiore a una inferiore. Poiché $\sum_{n=-\infty}^{\infty} |h(n)|$

è la massima uscita ottenibile con un ingresso che non supera mai l'unità, essa deve essere maggiore della risposta del filtro del secondo ordine ad una sinusoide di ampiezza unitaria alla frequenza di risonanza. Sulla base di questa osservazione si ottiene

$$\left(\sum_{n=0}^{\infty} |h(n)| \right)^2 \geq \frac{1}{(1-r)^2(1+r^2-2r \cos 2\theta)} \quad (9.31)$$

essendo il secondo membro di questa disuguaglianza il guadagno del filtro alla risonanza ($\omega = \theta$). Inoltre

$$\left(\frac{1}{\sin \theta} \sum_{n=0}^{\infty} r^n |\sin[(n+1)\theta]| \right)^2 \leq \left(\frac{1}{\sin \theta} \sum_{n=0}^{\infty} r^n \right)^2 \quad (9.32)$$

Perciò, in definitiva, per il filtro del secondo ordine

$$\frac{1}{2} \cdot 2^{-2b} \frac{1}{(1-r)^2(1+r^2-2r \cos 2\theta)} \leq \frac{\sigma_f^2}{\sigma_y^2} \leq \frac{1}{2} \cdot 2^{-2b} \frac{1}{\sin^2 \theta (1-r)^2} \quad (9.33)$$

I filtri molto selettivi in frequenza spesso richiedono poli prossimi al circolo unitario. Poiché le realizzazioni di tali filtri, in cascata o in parallelo, richiedono sistemi del primo e del secondo ordine, è importante esaminare le espressioni di sopra del rapporto rumore-segnale allorché i poli tendono a situarsi sul circolo unitario.

Per il filtro del primo ordine poniamo $\delta = 1 - |\alpha|$, così che per $\delta \rightarrow 0$ il polo tende al circolo unitario. Allora, in termini di δ , il rapporto rumore-segnale per il filtro del primo ordine è

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{4} \cdot 2^{-2b} \frac{1}{\delta^2} \quad (9.34)$$

Per il filtro del secondo ordine poniamo $\delta = 1 - r$ così che, ancora, per $\delta \rightarrow 0$, i poli tendono al circolo unitario. Allora, ammettendo che sia $\delta \ll 1$, si può approssimare $(1 + r^2 - 2r \cos 2\theta)$ con

$$1 + r^2 - 2r \cos 2\theta \cong 4 \sin^2 \theta + \delta^2 \quad (9.35)$$

che è approssimabile a $4 \sin^2 \theta$ se questo è grande rispetto a δ^2 . Di conseguenza, adottando questa approssimazione, si ha

$$\frac{1}{2} \cdot 2^{-2b} \frac{1}{4\delta^2 \sin^2 \theta} \leq \frac{\sigma_f^2}{\sigma_y^2} \leq \frac{1}{2} \cdot 2^{-2b} \frac{1}{\delta^2 \sin^2 \theta} \quad (9.36)$$

Osserviamo che il rapporto rumore-segnale per gli esempi fatti sopra può essere considerato proporzionale a $2^{-2b}/\delta^2$. Da ciò deriva che se δ è dimezzato, allora, per mantenere lo stesso rapporto rumore-segnale, b deve essere incrementato di uno: occorre, cioè, aggiungere un bit alla lunghezza di registro.

Nell'analisi precedente si è fatta l'ipotesi che l'ingresso del filtro sia rumore bianco uniformemente distribuito. Allorché δ tende a zero, la risposta in frequenza dei filtri sia del primo che del secondo ordine diventa più selettiva, così che una quota sempre maggiore dell'energia di ingresso viene

a trovarsi fuori banda. Una base alternativa per determinare il rapporto rumore-segnale è pertanto quella di usare un ingresso sinusoidale. Per questa scelta di ingressi non dovremo usare, ovviamente, per evitare la saturazione, la condizione generale (9.26), in quanto è possibile ora determinare esattamente la massima ampiezza di ingresso consentita come funzione dei parametri del filtro.

In particolare, se l'ingresso è della forma $x(n) = A \cos n\omega_0$, allora a regime l'uscita è della forma $y(n) = B \cos(n\omega_0 + \phi)$. Per evitare la saturazione B deve essere minore di uno, e per massimizzare l'energia del segnale di uscita B deve essere il più grande possibile. Pertanto il massimo rapporto segnale-rumore si ha quando A è scelto in modo che sia $y(n) = \cos(n\omega_0 + \phi)$. Si osservi che per poter scegliere A in questo modo la frequenza del segnale di ingresso deve essere nota. Per un ingresso sinusoidale di frequenza incognita, A deve essere scelto in modo che non si verifichi saturazione anche nel caso peggiore, cioè quando la frequenza dell'ingresso coincide con il picco di guadagno nella funzione di trasferimento del filtro [13].

Per i filtri in virgola fissa, nelle condizioni di validità del modello che abbiamo usato per l'errore di arrotondamento, il rumore di uscita è indipendente dalla frequenza e dall'ampiezza del segnale di ingresso. Pertanto, per questa scelta di ingressi, il rapporto rumore-segnale ottenuto per un filtro del primo ordine è

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{2^4} \cdot 2^{-2b} \frac{1}{1 - |\alpha|^2} \quad (9.37)$$

Se, come prima, poniamo $\alpha = 1 - \delta$, allora, per $\delta \ll 1$, si ha

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{48} \frac{2^{-2b}}{\delta} \quad (9.38)$$

In questo caso il rapporto rumore-segnale è proporzionale a $1/\delta$ invece che a $1/\delta^2$, così che, se δ viene moltiplicato per $1/4$ e la lunghezza di registro aumentata di un bit, il rapporto rumore-segnale rimane costante. Il caso dei filtri del secondo ordine può essere considerato in modo analogo. Anche ora per un ingresso sinusoidale l'uscita con ampiezza massima avrà la forma $y(n) = \cos(n\omega_0 + \phi)$, e il rapporto rumore-segnale in questo caso sarà

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{1^2} \cdot 2^{-2b} \frac{1 + r^2}{1 - r^2} \frac{1}{1 + r^4 - 2r^2 \cos 2\theta} \quad (9.39)$$

Ponendo anche in questo caso $r = 1 - \delta$, per $\delta \ll 1$, avremo

$$\frac{\sigma_f^2}{\sigma_y^2} \cong \frac{2^{-2b}}{4\delta \sin^2 \theta} \quad (9.40)$$

Come per i filtri del primo ordine, il rapporto rumore-segnale è proporzionale a $1/\delta$ anziché a $1/\delta^2$. Il confronto tra i rapporti rumore-segnale per

un ingresso rumore bianco e un ingresso sinusoidale serve ad illustrare le conseguenze della scelta di un particolare tipo di ingresso. In un certo senso i due casi considerati rappresentano casi estremi. Quando l'ingresso risulta limitato a una stretta e conosciuta banda di frequenze, allora l'analisi precedente relativa a un ingresso sinusoidale è più rappresentativa, mentre, quando l'ingresso risulta più a larga banda, è più rappresentativa l'analisi con in ingresso un rumore bianco.

Nella discussione precedente, il rapporto rumore-segnale per il caso di ingresso rumore bianco è stato derivato assumendo che l'ampiezza dell'ingresso fosse abbastanza piccola da evitare la saturazione nel caso più generale. Nei casi pratici un cambiamento di scala dell'ingresso sulla base della (9.26) può considerarsi pessimistico, in quanto la probabilità che nella (9.25) si raggiunga il segno di uguaglianza è estremamente piccola. In pratica si consente all'ampiezza di essere alquanto maggiore di quanto indicato dalla (9.26) e, se il risultato di un'addizione dovesse provocare saturazione, si usa bloccare l'uscita al massimo valore nello stesso modo in cui nel par. 9.2 abbiamo assunto che venissero bloccati i campioni di ingresso. Questo approccio è comunemente chiamato *aritmetica con saturazione*. Nella realizzazione del filtro la saturazione rappresenta ovviamente una distorsione, e la scelta del fattore di scala dell'ingresso dipenderà da quanto spesso ci si può permettere tale distorsione.

Il tipo di analisi illustrato nella discussione precedente può essere applicato per indagare gli effetti della quantizzazione in sistemi definiti da equazioni lineari alle differenze di ordine più elevato. È tuttavia difficile ottenere risultati che siano di vasta applicazione. Ciò avviene perché gli effetti della quantizzazione dipendono fortemente dalle proprietà della funzione di trasferimento che si desidera e dalla struttura specifica usata per realizzare quella funzione di trasferimento.

Come esempio, si considerino le due realizzazioni di un sistema del secondo ordine illustrate nella fig. 9.9. Le sorgenti di rumore nei due sistemi si manifestano all'uscita in due modi diversi. Pertanto i due sistemi avranno in generale all'uscita diversi rapporti rumore-segnale per lo stesso segnale di ingresso. Non è possibile tuttavia stabilire quale delle due forme è da preferirsi a meno che non si conoscano i parametri.

Per sistemi di ordine più elevato, la struttura di tipo parallelo è la più semplice da analizzare. In questo caso, per determinare il rapporto rumore-segnale all'uscita, è sufficiente un'analisi dei sistemi del primo e del secondo ordine (completi di zeri), in quanto gli effetti della quantizzazione si assumono indipendenti fra sezione e sezione. Tuttavia, anche in questo caso il rapporto rumore-segnale dipende dalla struttura delle sezioni del secondo ordine. La struttura in cascata presenta problemi ancora più difficili, in quanto l'ordine in cui sono sistemati poli e zeri può avere grande effetto sul rapporto rumore-segnale complessivo, per il fatto che il rumore generato in una particolare sezione del secondo ordine è filtrato da tutte le sezioni successive. Qui nasce allora l'interessante problema di determi-

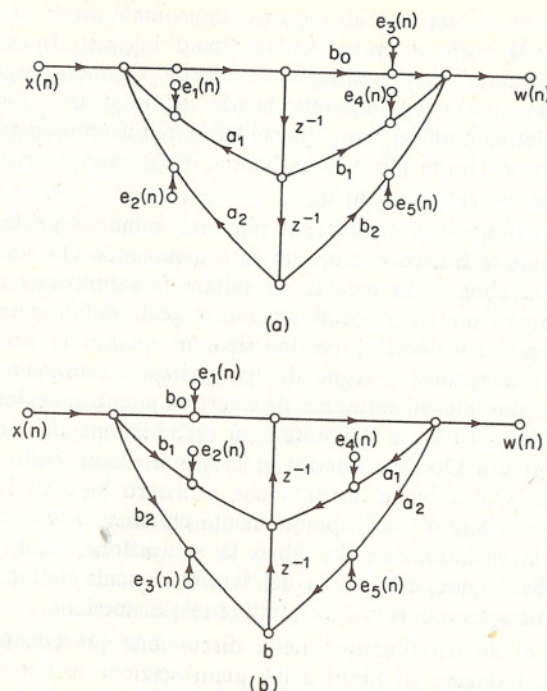


Fig. 9.9 Modelli statistici per il rumore di arrotondamento in sistemi IIR del secondo ordine: (a) poli prima degli zeri; (b) zeri prima dei poli.

Il migliore accoppiamento di zeri e poli e il migliore ordinamento delle risultanti sezioni del secondo ordine al fine di minimizzare il rapporto rumore-segnale in uscita. Questo problema è complicato dal fatto che i segnali devono essere cambiati di scala in modo che non si verifichi saturazione in nessun punto della catena dei sistemi del secondo ordine. Un'analisi dettagliata degli effetti della quantizzazione nelle realizzazioni in cascata e in parallelo è fornita da Jackson [13, 14]. L'indicazione che si ricava da questo lavoro è che per la struttura parallelo è piccola la dipendenza dalla forma usata, fig. 9.9(a) o (b), per realizzare le sezioni del secondo ordine. Invece, per la struttura in cascata, esiste una maggiore dipendenza dalla forma delle sezioni del secondo ordine. Sembra infine che la struttura in parallelo sia leggermente migliore della struttura in cascata, sia pure realizzata con il più conveniente ordinamento delle sezioni.

9.3.3 Analisi statistica della quantizzazione nelle realizzazioni in virgola mobile di filtri numerici IIR

Dalla discussione precedente risulta chiaro che la ristretta dinamica dell'aritmetica in virgola fissa rende necessaria un'accurata correzione di scala per gli ingressi e per i livelli di segnale intermedi in tutte le realizza-

zioni di filtri numerici in virgola fissa. La necessità di tali correzioni di scala può essere sostanzialmente eliminata usando l'aritmetica in virgola mobile. Ciò implica tuttavia che si introduce del rumore sia per le moltiplicazioni che per le addizioni [15-18]. Come si è visto nel par. 9.1, se $Q[x(n)]$ indica il risultato dell'arrotondamento applicato alla mantissa di un segnale rappresentato in virgola mobile, allora $Q[x(n)]$ si può esprimere come

$$Q[x(n)] = x(n)(1 + \varepsilon(n)) = x(n) + x(n)\varepsilon(n) \quad (9.41)$$

dove, con la lunghezza della mantissa pari a $(b + 1)$ bit, l'errore relativo soddisfa la relazione

$$-2^{-b} < \varepsilon(n) \leq 2^{-b} \quad (9.42)$$

Pertanto, per un filtro realizzato con l'aritmetica in virgola mobile, possiamo rappresentare l'effetto della quantizzazione con un termine di errore additivo $e(n) = x(n)\varepsilon(n)$. Anche ora, consideriamo il filtro del primo ordine rappresentato nella fig. 9.10. La fig. 9.10(a) mostra il sistema del primo ordine nell'ipotesi di una aritmetica con precisione infinita. Pertanto $y(n)$ è l'uscita esatta corrispondente all'ingresso $x(n)$. La fig. 9.10(b) rappresenta il sistema quando vengono inseriti dei quantizzatori in virgola mobile dopo la moltiplicazione e l'addizione per tener conto dell'arrotondamento della mantissa. In questo caso $w(n)$ rappresenta l'uscita rumorosa dovuta all'ingresso $x(n)$. Nella fig. 9.10(c) i quantizzatori sono sostituiti da sorgenti di rumore additivo corrispondenti all'arrotondamento in virgola mobile. È importante osservare che, se $e_1(n)$ ed $e_2(n)$ sono note, la fig. 9.10(b) è equivalente alla fig. 9.10(c). Per applicare l'analisi statistica dobbiamo fare alcune ipotesi sulle sorgenti di rumore $e_1(n)$ ed $e_2(n)$. Di primo acchito tali ipotesi possono sembrare difficili da giustificare. Tuttavia, come sottolineato in precedenza, noi non stiamo tentando di sviluppare un'analisi molto precisa. Entro limiti ristretti, i risultati basati sul modello che stiamo per sviluppare e usare sono stati verificati sperimentalmente.

Cominciamo con l'assumere che $x(n)$, l'ingresso, sia un processo casuale, così che l'ingresso e l'uscita del filtro possano essere descritti in termini di medie. Per convenienza assumiamo che $x(n)$ abbia media nulla. Osserviamo ora che le sorgenti di errore $e_1(n)$ ed $e_2(n)$ sono date da

$$e_1(n) = \varepsilon_1(n)aw(n-1) \quad (9.43)$$

e

$$e_2(n) = \varepsilon_2(n)g(n) \quad (9.44)$$

Senza quantizzazione è $w(n-1) = y(n-1)$ e $g(n) = y(n)$, così che, se gli errori sono piccoli, possiamo approssimare $e_1(n)$ ed $e_2(n)$ con

$$e_1(n) \approx a\varepsilon_1(n)y(n-1) \quad (9.45)$$

$$e_2(n) \approx \varepsilon_2(n)y(n) \quad (9.46)$$

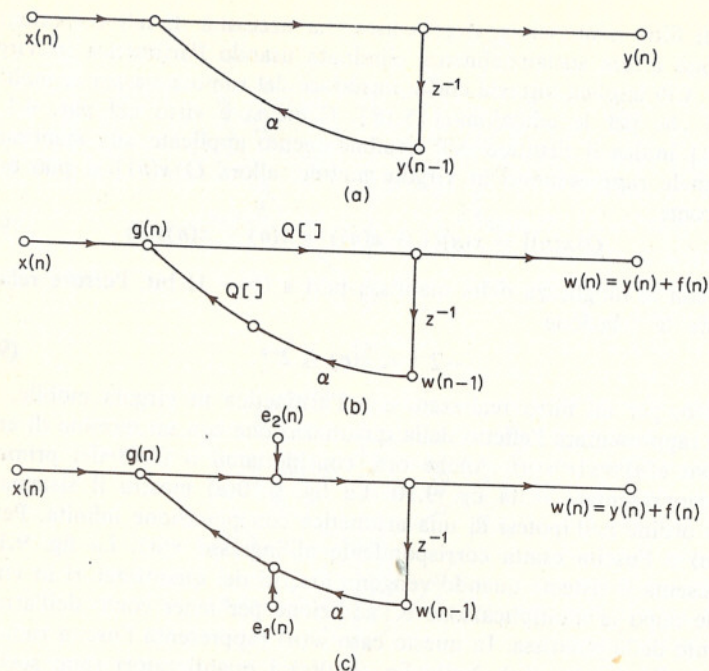


Fig. 9.10 Sistemi IIR del primo ordine: (a) sistema lineare ideale; (b) modello non lineare; (c) modello statistico per rumore di arrotondamento in virgola mobile.

La conseguenza di questa ulteriore approssimazione è di poter esprimere l'errore additivo in termini dei segnali nel filtro ideale non quantizzato invece che nel filtro effettivo. Oltre a tale approssimazione, noi assumiamo, come prima, che gli errori relativi $\varepsilon_1(n)$ ed $\varepsilon_2(n)$ siano

1. Rumore bianco.
2. Scorrelati l'uno con l'altro.
3. Scorrelati con l'ingresso od ogni altra variabile di nodo del sistema.
4. Distribuiti uniformemente in ampiezza nel campo di valori da -2^{-b} a $+2^{-b}$ [v. (9.8 a)].

Poiché $\varepsilon_1(n)$ è un rumore bianco ed è scorrelato con $y(n-1)$, ne segue (v. probl. 4 del cap. 8) che $e_1(n)$ è bianco con varianza

$$\sigma_{e_1}^2 = \alpha^2 \sigma_{e_1}^2 \cdot E[y^2(n-1)] \quad (9.47)$$

od anche, assumendo che $x(n)$, e conseguentemente $y(n)$, abbia media nulla,

$$\sigma_{e_1}^2 = \alpha^2 \sigma_{e_1}^2 \cdot \sigma_y^2 \quad (9.48)$$

Analogamente $e_2(n)$ è bianco con varianza

$$\sigma_{e_2}^2 = \sigma_{e_2}^2 \sigma_y^2 \quad (9.49)$$

Indicando con $h_1(n)$ ed $h_2(n)$ le risposte all'impulso dagli ingressi delle sorgenti di rumore all'uscita, e con $f_1(n)$ ed $f_2(n)$ le componenti dell'errore in uscita dovute a $e_1(n)$ ed $e_2(n)$, si ha

$$f(n) = f_1(n) + f_2(n)$$

Avendo assunto $\varepsilon_1(n)$ ed $\varepsilon_2(n)$ scorrelati, ne segue che anche $e_1(n)$ ed $e_2(n)$ sono scorrelati, e lo sono anche le corrispondenti uscite di rumore $f_1(n)$ ed $f_2(n)$. Pertanto la varianza del rumore di uscita è

$$\sigma_f^2 = \sigma_{f_1}^2 + \sigma_{f_2}^2 \quad (9.50)$$

dove

$$\sigma_{f_1}^2 = \sigma_{e_1}^2 \sum_{n=-\infty}^{\infty} h_1^2(n) \quad (9.51a)$$

e

$$\sigma_{f_2}^2 = \sigma_{e_2}^2 \sum_{n=-\infty}^{\infty} h_2^2(n) \quad (9.51b)$$

Poiché

$$h_1(n) = h_2(n) = \alpha^n u(n) \quad (9.52)$$

dalle (9.48)-(9.52) segue che

$$\sigma_f^2 = \sigma_y^2 \frac{1}{1 - \alpha^2} (\alpha^2 \sigma_{e_1}^2 + \sigma_{e_2}^2) \quad (9.53)$$

Infine, poiché si è assunto che $\varepsilon_1(n)$ ed $\varepsilon_2(n)$ abbiano entrambi densità di probabilità uniforme tra -2^{-b} e $+2^{-b}$, è

$$\sigma_{e_1}^2 = \sigma_{e_2}^2 = \frac{1}{3} \cdot 2^{-2b}$$

Per cui

$$\sigma_f^2 = \frac{1}{3} \cdot 2^{-2b} \sigma_y^2 \frac{1 + \alpha^2}{1 - \alpha^2} \quad (9.54)$$

e il rapporto rumore-segnale in uscita è

$$\frac{\sigma_f^2}{\sigma_y^2} = \frac{1}{3} \cdot 2^{-2b} \frac{1 + \alpha^2}{1 - \alpha^2} \quad (9.55)$$

È interessante osservare che il rapporto rumore-segnale espresso dalla (9.55) per i filtri del primo ordine è stato derivato senza fare particolari ipotesi sulle proprietà spettrali dell'ingresso. Perciò la (9.55) si applica sia per un ingresso a larga banda, come un rumore bianco, che a banda stretta, come un ingresso sinusoidale. Per filtri di tipo più generale, tuttavia, il rapporto rumore-segnale dipenderà dalla forma dell'ingresso.

In modo simile a quanto si è appena visto, possiamo analizzare un filtro del secondo ordine. La fig. 9.11(a) rappresenta il filtro ideale del secondo ordine mentre la fig. 9.11(b) rappresenta il filtro con incluse le sorgenti di rumore. Si osservi che, dovendosi includere anche le sorgenti

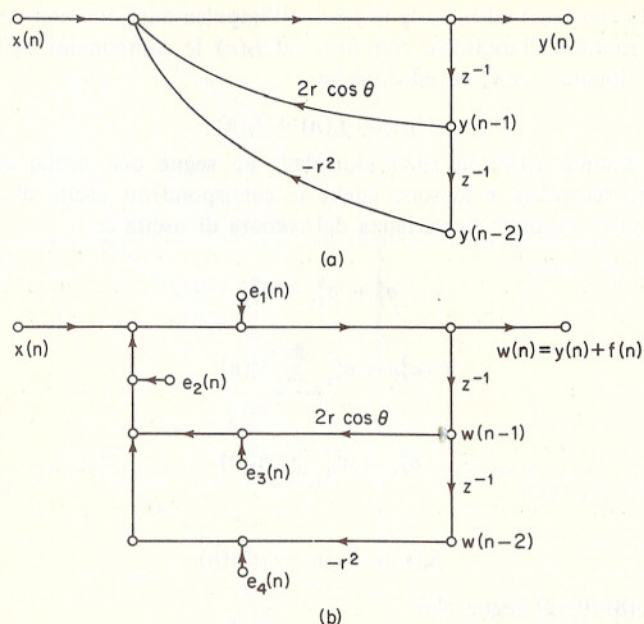


Fig. 9.11 Sistema IIR del secondo ordine: (a) sistema lineare ideale; (b) modello statistico per rumore di arrotondamento in virgola mobile

di rumore dovuto alle addizioni, l'ordine nel quale vengono sommati i prodotti è importante. La fig. 9.11(b) corrisponde al caso in cui si sommano per primi i prodotti (arrotondati) $2r \cos \theta w(n-1)$ e $-r^2 w(n-2)$ e poi si somma l'ingresso $x(n)$ alla somma arrotondata. Le sorgenti di rumore $e_3(n)$ ed $e_4(n)$ rappresentano il rumore dovuto alle moltiplicazioni, e le sorgenti di rumore $e_1(n)$ ed $e_2(n)$ rappresentano il rumore dovuto alle addizioni. Con ipotesi simili a quelle di prima, nelle quali abbiamo trascurato i termini del secondo ordine, scriviamo che

$$\begin{aligned} e_1(n) &= y(n)e_1(n) \\ e_2(n) &= (y(n) - x(n))e_2(n) \\ e_3(n) &= 2r \cos \theta y(n-1)e_3(n) \\ e_4(n) &= -r^2 y(n-2)e_4(n) \end{aligned} \quad (9.56)$$

Ancora, assumiamo che $e_1(n)$, $e_2(n)$, $e_3(n)$, ed $e_4(n)$ siano scorrelati l'uno con l'altro e con $x(n)$, siano sequenze rumore bianco, ed abbiano funzioni di densità di probabilità uniformi. Poiché $e_1(n)$, $e_2(n)$, $e_3(n)$ ed $e_4(n)$ sono bianche e scorrelate con $x(n)$ [e conseguentemente anche con $y(n)$], ne discende, come nell'esempio precedente, che $e_1(n)$, $e_2(n)$, $e_3(n)$ ed $e_4(n)$

sono bianche e le loro varianze sono date da

$$\begin{aligned} \sigma_{e_1}^2 &= E[y^2(n)] \cdot \sigma_{e_1}^2 \\ \sigma_{e_2}^2 &= E[(y(n) - x(n))^2] \cdot \sigma_{e_2}^2 \\ \sigma_{e_3}^2 &= 4r^2 \cos^2 \theta E[y^2(n-1)] \cdot \sigma_{e_3}^2 \\ \sigma_{e_4}^2 &= r^4 E[y^2(n-2)] \cdot \sigma_{e_4}^2 \end{aligned} \quad (9.57)$$

Poiché ognuna delle quattro sorgenti di rumore è per ipotesi scorrelata, la varianza della sequenza di rumore di uscita, $f(n)$, è la somma delle varianze dovute a ciascuna delle sorgenti di rumore di arrotondamento. Pertanto la varianza del rumore di uscita è

$$\sigma_f^2 = \sigma_{e_1}^2 \sum_{n=-\infty}^{+\infty} h_1^2(n) + \sigma_{e_2}^2 \sum_{n=-\infty}^{+\infty} h_2^2(n) + \sigma_{e_3}^2 \sum_{n=-\infty}^{+\infty} h_3^2(n) + \sigma_{e_4}^2 \sum_{n=-\infty}^{+\infty} h_4^2(n)$$

Dalla fig. 9.11 osserviamo che

$$h_1(n) = h_2(n) = h_3(n) = h_4(n) = \frac{1}{\sin \theta} r^2 \sin(n+1)\theta u(n) \quad (9.58)$$

e quindi

$$\begin{aligned} \sum_{n=-\infty}^{+\infty} h_1^2(n) &= \sum_{n=-\infty}^{+\infty} h_2^2(n) = \sum_{n=-\infty}^{+\infty} h_3^2(n) = \sum_{n=-\infty}^{+\infty} h_4^2(n) \\ &= \frac{1+r^2}{1-r^2 r^4 + 1 - 2r^2 \cos 2\theta} \quad (9.59) \end{aligned}$$

Se indichiamo con G il termine di destra della (9.59), è

$$\sigma_f^2 = G(\sigma_{e_1}^2 + \sigma_{e_2}^2 + \sigma_{e_3}^2 + \sigma_{e_4}^2)$$

Tenendo poi conto del fatto che

1. $\sigma_{e_1}^2 = \sigma_{e_2}^2 = \sigma_{e_3}^2 = \sigma_{e_4}^2 = \frac{1}{3} \cdot 2^{-2b}$
2. $E[y^2(n)] = E[y^2(n-1)] = E[y^2(n-2)] = \sigma_y^2$
3. $E[(y(n) - x(n))^2] = E[y^2(n)] + E[x^2(n)] - 2E[y(n)x(n)]$
 $= \sigma_y^2 + \sigma_x^2 - 2E[y(n)x(n)]$

possiamo scrivere

$$\sigma_f^2 = \frac{1}{3} \cdot 2^{-2b} G \sigma_y^2 (2 + r^4 + 4r^2 \cos^2 \theta) + \frac{1}{3} \cdot 2^{-2b} G (\sigma_x^2 - 2E[y(n)x(n)]) \quad (9.60)$$

Senza fare altre ipotesi la (9.60) non può essere ulteriormente semplificata. Tuttavia, se assumiamo che l'ingresso sia bianco, allora

$$\sigma_y^2 = \sigma_x^2 \sum_{n=-\infty}^{+\infty} h^2(n) = G \sigma_x^2 \quad (9.61)$$

Infine osserviamo che (v. probl. 9 del cap. 8) $E[y(n)x(n)] = h(0)\sigma_x^2$, oppure, essendo $h(0) = 1$,

$$E[y(n)x(n)] = \sigma_x^2 \quad (9.62)$$

Con queste espressioni la (9.60) può essere riscritta come

$$\begin{aligned}\sigma_f^2 &= \frac{1}{3} \cdot 2^{-2b} G \sigma_y^2 (2 + r^4 + 4r^2 \cos^2 \theta) - \frac{1}{3} \cdot 2^{-2b} G \sigma_x^2 \\ &= \frac{1}{3} \cdot 2^{-2b} \sigma_y^2 \{ (2 + r^4 + 4r^2 \cos^2 \theta) G - 1 \}\end{aligned}\quad (9.63)$$

Nel caso in cui il guadagno è elevato, diventa possibile confrontare l'aritmetica in virgola fissa e quella in virgola mobile approssimando le espressioni dei rapporti rumore-segnale. Per il filtro del primo ordine, con $\alpha = 1 - \delta$ e $|\delta| \ll 1$, la (9.55) valida per l'aritmetica in virgola mobile è approssimabile con

$$\frac{\sigma_f^2}{\sigma_y^2} \simeq \frac{1}{3} \cdot 2^{-2b} \frac{1}{\delta} \quad (9.64)$$

Analogamente, per il filtro del secondo ordine, con $r = 1 - \delta$ e $\delta \ll 1$, si ha

$$\frac{\sigma_f^2}{\sigma_y^2} \simeq \frac{1}{3} \cdot 2^{-2b} \frac{3 + 4 \cos^2 \theta}{4\delta \sin^2 \theta} \quad (9.65)$$

Per l'aritmetica in virgola fissa ricordiamo che, con un rumore bianco come ingresso, il rapporto rumore-segnale si comporta come $1/\delta^2$, e come $1/\delta$ se l'ingresso è una sinusoide. Il confronto delle (9.64) e (9.65) con le (9.34) e (9.36) indica un rapporto rumore-segnale significativamente più piccolo per l'aritmetica in virgola mobile rispetto all'aritmetica in virgola fissa nel caso in cui l'ingresso è un rumore bianco. È importante tener presente che per i filtri in virgola fissa i rapporti rumore-segnale sono stati calcolati nell'ipotesi che i segnali di ingresso siano i più grandi possibili. Se il livello del segnale di ingresso diminuisce, il rapporto rumore-segnale aumenta, in quanto la varianza del rumore di uscita è indipendente dal livello del segnale di ingresso. Per l'aritmetica in virgola mobile, d'altra parte, la varianza del rumore di uscita è proporzionale alla varianza del segnale di uscita, e pertanto, se il livello dell'ingresso è modificato in su o in giù, la stessa cosa avviene per il rumore di arrotondamento. È anche importante osservare che, nel confronto appena discusso, è implicita l'ipotesi che la mantissa in virgola mobile sia uguale in lunghezza all'intera parola in virgola fissa e non si tiene pertanto conto dei bit in più necessari per la caratteristica.

Come per il caso della virgola fissa, l'analisi dei sistemi di ordine più elevato diventa molto complicata. Il rumore di quantizzazione dipende dai parametri del sistema e dalla struttura usata per realizzarlo. In generale, i commenti fatti alla fine del par. 9.3.2 si applicano anche alle realizzazioni dei filtri numerici in virgola mobile. Tuttavia, poiché con l'aritmetica in virgola mobile non si hanno praticamente problemi di dinamica di segnale, le realizzazioni con struttura in cascata non sono probabilmente troppo sensibili all'ordinamento dei poli e degli zeri.

9.4 EFFETTI DELLA LUNGHEZZA FINITA DEI REGISTRI NELLE REALIZZAZIONI DI FILTRI NUMERICI FIR

Il tipo di analisi sviluppato nel paragrafo precedente può anche essere applicato allo studio degli effetti della quantizzazione nei filtri numerici FIR. Per certi riguardi questa analisi è più semplice di quella dei filtri IIR. Per esempio, non ci sono effetti di ciclo limite quando si usano realizzazioni non ricorsive, come quelle in forma diretta o in cascata, in quanto tali strutture non hanno reazione. Se la risposta all'impulso è lunga N campioni, l'uscita di un filtro FIR realizzato in forma diretta o in cascata deve annullarsi dopo che l'ingresso è stato zero per N campioni consecutivi. Al contrario, le realizzazioni ricorsive dei sistemi FIR, quali la struttura a campionamento in frequenza, sono soggette ai problemi discussi nel par. 9.3, e in particolare l'analisi dei sistemi del secondo ordine si applica alla forma del campionamento in frequenza [19].

La dinamica del segnale e il rumore di arrotondamento sono importanti per i sistemi FIR proprio come lo sono per i sistemi IIR. In questo paragrafo esamineremo alcuni dei problemi che ne derivano per realizzazioni in virgola fissa e in virgola mobile, sia per la forma diretta che per la forma in cascata.

9.4.1 Analisi statistica della quantizzazione nelle realizzazioni in virgola fissa di filtri numerici FIR

Si consideri un filtro lineare invariante alla traslazione con risposta all'impulso $h(n)$, diversa da zero solo per $0 \leq n \leq N-1$. La realizzazione in forma diretta di tale sistema è il calcolo diretto della somma di convoluzione

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k)$$

Il grafo di flusso per la realizzazione in forma diretta è mostrato nella fig. 9.12(a). La fig. 9.12(b) mostra la stessa struttura con in più le sorgenti di rumore sommate in modo da render conto degli effetti di arrotondamento dei prodotti $h(k)x(n-k)$. È applicato inoltre all'ingresso un guadagno costante A allo scopo di evitare la saturazione. Come prima, assumiamo che:

1. Le sorgenti $e_k(n)$ siano rumori bianchi.
2. Gli errori siano uniformemente distribuiti su un intervallo di quantizzazione.
3. Le sorgenti di errore siano scorrelate l'una con l'altra e con l'ingresso.

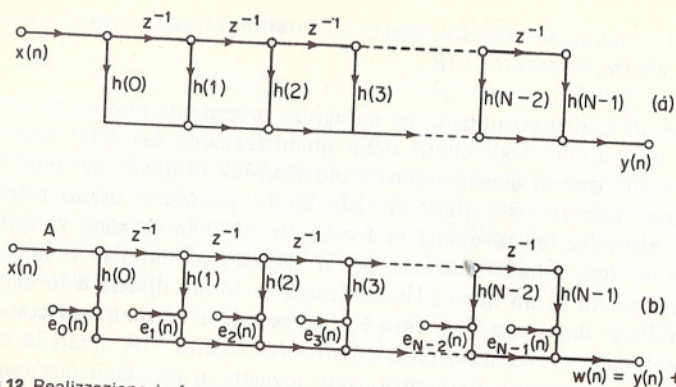


Fig. 9.12 Realizzazione in forma diretta di un sistema FIR: (a) sistema lineare ideale; (b) modello statistico per rumore di arrotondamento in virgola fissa

Poniamo innanzitutto $A = 1$. Dalla fig. 9.12(b) risulta chiaro che ogni sorgente di rumore si somma direttamente all'uscita, e perciò il rumore di uscita è

$$f(n) = \sum_{k=0}^{N-1} e_k(n)$$

Poiché le sorgenti di rumore sono per ipotesi indipendenti, la varianza del rumore di uscita è (per l'arrotondamento)

$$\sigma_f^2 = N \frac{2^{-2b}}{12}$$

e la media è zero. Si osservi che per la forma diretta, il livello di rumore in uscita è indipendente dai parametri del filtro in quanto il rumore non viene elaborato affatto dal sistema. Si noti ancora che il rumore è proporzionale ad N , la lunghezza della sequenza risposta al campione unitario.

La limitazione di dinamica propria dell'aritmetica in virgola fissa impone variazioni di scala in ingresso in modo che non si abbia saturazione. Abbiamo visto [v. la (9.25)] che una limitazione superiore sull'uscita di un sistema lineare invariante alla traslazione è

$$|y(n)| \leq x_{\max} \sum_{n=-\infty}^{\infty} |h(n)|$$

dove x_{\max} è la massima ampiezza del segnale di ingresso. Per garantire che non si abbia saturazione deve essere $|y(n)| < 1$ per tutti gli n . Ciò implica che il guadagno in ingresso deve soddisfare la disuguaglianza

$$A < \frac{1}{x_{\max} \sum_{n=0}^{N-1} |h(n)|} \quad (9.66)$$

Tale correzione di scala risulta appropriata per un segnale a larga banda, come il rumore bianco, ma sarebbe troppo cautelativa per un segnale a

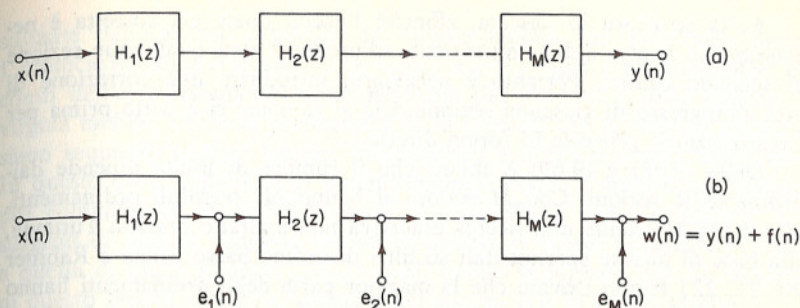


Fig. 9.13 Realizzazione in cascata di un sistema FIR: (a) sistema lineare ideale; (b) modello statistico per rumore di arrotondamento in virgola fissa

banda stretta, quale una sinusoidale. In quest'ultimo caso abbiamo visto che l'ingresso dovrebbe essere modificato in relazione al picco della risposta in frequenza del sistema. Pertanto una scelta alternativa del coefficiente di scala dell'ingresso è

$$A < \frac{1}{x_{\max} \cdot \max_{0 \leq \omega \leq \pi} [|H(e^{j\omega})|]} \quad (9.67)$$

Occorre osservare (v. probl. 2 di questo capitolo) che con l'aritmetica in complemento a due, quando vengono sommati insieme, come in fig. 9.12, più di due numeri, se il risultato ideale della somma è minore di uno, allora possono verificarsi saturazioni nel calcolo delle somme parziali, ottenendo tuttavia il risultato finale giusto. La correzione di scala indicata dalla (9.66) garantirà comunque uscite corrette per le realizzazioni dei sistemi in forma diretta (eccetto, ovviamente, che per il rumore di arrotondamento).

Un filtro numerico FIR può anche essere realizzato con una cascata di sezioni del secondo ordine come nella fig. 9.13(a), dove ogni sezione del secondo ordine $H_k(z)$ è realizzata in forma diretta come in fig. 9.12(a). Assumiamo per comodità che N sia dispari per cui $M = (N-1)/2$. Poiché ogni sezione del secondo ordine ha in uscita tre sorgenti di rumore bianco indipendenti, gli effetti della quantizzazione possono essere rappresentati come nella fig. 9.13(b), dove ogni sorgente di rumore $e_k(n)$ ha varianza $3(2^{-2b}/12) = 2^{-2b}/4$. In tal caso una data sorgente di rumore $e_k(n)$ è filtrata dalle sezioni successive, così che la varianza del rumore in uscita dipenderà dall'ordine delle sezioni del secondo ordine nella catena. Se indichiamo con $g_i(n)$ la risposta all'impulso dalla sorgente di rumore $e_i(n)$ all'uscita, possiamo scrivere

$$\sigma_{e_i}^2 = \frac{2^{-2b}}{4} \left(\sum_{n=0}^{N-2i} g_i^2(n) \right) \quad (9.68)$$

e la varianza del rumore totale in uscita è

$$\sigma_f^2 = \sum_{i=1}^M \sigma_{e_i}^2 = \frac{2^{-2b}}{4} \left(\sum_{i=1}^M \sum_{n=0}^{N-2i} g_i^2(n) \right) \quad (9.69)$$

Nella struttura in cascata, affinché l'uscita finale sia corretta è necessario che non si abbia saturazione all'uscita di una qualunque sezione del secondo ordine. Pertanto è necessario introdurre una correzione di scala all'ingresso di ciascuna sezione. Ciò si fa come si è detto prima per la realizzazione generale in forma diretta.

Dalle (9.68) e (9.69) è chiaro che il rumore di uscita dipende dall'ordine delle sezioni. Con M sezioni si hanno $M!$ possibili ordinamenti, per cui per M grande una ricerca esaustiva non è affatto pratica. Tuttavia, sulla base di misure sperimentali su filtri di ordine basso, Chan e Rabiner [20, 21, 22] hanno trovato che la maggior parte degli ordinamenti hanno proprietà di rumore ragionevolmente buone e hanno fornito un algoritmo di ricerca di un buon ordinamento. I loro risultati indicano che un buon ordinamento è uno per il quale la risposta in frequenza da ogni sorgente di rumore all'uscita è relativamente piatta e dotata di un basso picco di guadagno.

9.4.2 Analisi statistica degli effetti della quantizzazione nelle realizzazioni in virgola mobile di filtri numerici FIR

L'uso dell'aritmetica in virgola mobile può eliminare di fatto ogni preoccupazione per la dinamica del segnale nella realizzazione dei filtri numerici FIR. È questo un approccio particolarmente conveniente quando un filtro va realizzato come un programma per un calcolatore di uso generale, dove l'attenzione è maggiormente rivolta alla facilità di programmazione piuttosto che alla velocità di calcolo. Tuttavia, spesso la maggiore complessità dell'aritmetica in virgola mobile non trova giustificazione nelle realizzazioni circuitali « special purpose » dove la maggiore preoccupazione è l'economicità.

Si consideri la realizzazione in forma diretta di un sistema FIR come rappresentato nella fig. 9.14(a). La fig. 9.14(b) rappresenta il sistema ri-

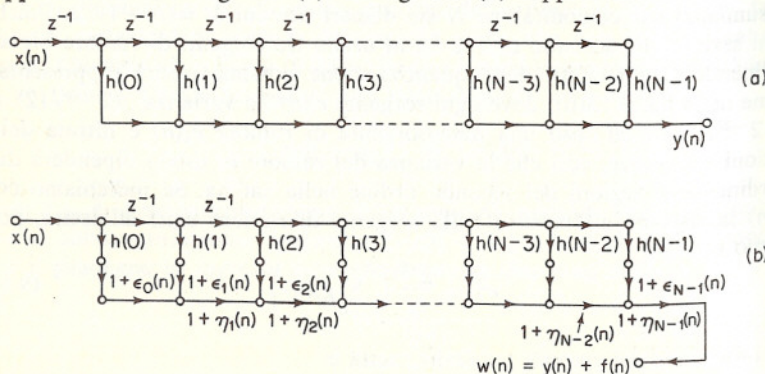


Fig. 9.14 Realizzazione in forma diretta di un sistema FIR: (a) sistema lineare ideale; (b) modello statistico per rumore di arrotondamento in virgola mobile

sultante dall'uso dell'aritmetica in virgola mobile con precisione finita. I coefficienti tempo-varianti $[1 + \epsilon_k(n)]$ e $[1 + \eta_k(n)]$ rappresentano gli effetti dell'arrotondamento rispettivamente dei prodotti e delle somme in virgola mobile. Come in precedenza, si fa l'ipotesi che gli $\epsilon_k(n)$ e gli $\eta_k(n)$ siano sequenze rumore bianco, indipendenti e uniformemente distribuite. In questo caso il modello di sistema tempo-variante comporta un insieme più compatto di equazioni rispetto a un modello tempo-invariante con sorgenti di rumore additivo. La nostra analisi in questo paragrafo è simile a quella di Liu e Kaneko [16], i quali hanno sviluppato con un modello di questo genere un'analisi generale dei sistemi IIR in forma diretta.

Dalla fig. 9.14 possiamo esprimere l'uscita $w(n)$ come

$$\begin{aligned} w(n) &= (1 + \epsilon_0(n)) \prod_{r=1}^{N-1} (1 + \eta_r(n)) h(0) x(n) \\ &+ (1 + \epsilon_1(n)) \prod_{r=1}^{N-1} (1 + \eta_r(n)) h(1) x(n-1) + \dots \\ &+ (1 + \epsilon_k(n)) \prod_{r=k}^{N-1} (1 + \eta_r(n)) h(k) x(n-k) + \dots \\ &+ (1 + \epsilon_{N-1}(n)) (1 + \eta_{N-1}(n)) h(N-1) x(n-N+1) \\ &= \sum_{k=0}^{N-1} A(n, k) h(k) x(n-k) \end{aligned} \quad (9.70a)$$

dove

$$A(n, 0) = (1 + \epsilon_0(n)) \prod_{r=1}^{N-1} (1 + \eta_r(n)) \quad (9.70b)$$

$$A(n, k) = (1 + \epsilon_k(n)) \prod_{r=k}^{N-1} (1 + \eta_r(n)), \quad k \neq 0 \quad (9.70c)$$

Se assumiamo che sia $w(n) = y(n) + f(n)$, allora dalla (9.70a) segue che

$$f(n) = \sum_{k=0}^{N-1} [A(n, k) - 1] h(k) x(n-k) \quad (9.71)$$

Si può far vedere che la quantità $[A(n, k) - 1]$ ha media nulla, così che il rumore di uscita ha media nulla. L'espressione generale per la varianza del rumore di uscita è perciò

$$\begin{aligned} \sigma_f^2 &= E[f^2(n)] \\ &= E \left[\sum_{k=0}^{N-1} \sum_{l=0}^{N-1} (A(n, k) - 1)(A(n, l) - 1) x(n-k) x(n-l) h(k) h(l) \right] \\ &= \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} E[(A(n, k) - 1)(A(n, l) - 1)] h(k) h(l) \phi_{xx}(l-k) \end{aligned} \quad (9.72)$$

Se l'ingresso è un segnale casuale con lo spettro di potenza piatto e varianza σ_x^2 , allora la (9.72) diventa

$$\begin{aligned}\sigma_f^2 &= \sigma_x^2 \sum_{k=0}^{N-1} h^2(k) E[(A(n, k) - 1)^2] \\ &= \sigma_x^2 \sum_{k=0}^{N-1} h^2(k) \{E[A^2(n, k)] - 1\}\end{aligned}\quad (9.73)$$

Dalle (9.70b) e (9.70c) e le ipotesi fatte sulle quantità $\varepsilon_k(n)$ e $\eta_k(n)$ ricaviamo

$$E[A^2(n, 0)] = \left(1 + \frac{2^{-2b}}{3}\right)^N \quad (9.74a)$$

$$E[A^2(n, k)] = \left(1 + \frac{2^{-2b}}{3}\right)^{N+1-k}, \quad k \neq 0 \quad (9.74b)$$

Osserviamo che usando la (9.74b) per $k = 0$ si commette solo un errore piuttosto piccolo. Inoltre, poiché $2^{-2b}/3 \ll 1$ in qualsiasi ragionevole situazione, è possibile esprimere la (9.74b) usando un'approssimazione binomiale come

$$E[A^2(n, k)] = 1 + (N + 1 - k) \frac{2^{-2b}}{3} \quad (9.75)$$

Pertanto σ_f^2 diventa

$$\sigma_f^2 = (N + 1) \frac{2^{-2b}}{3} \sigma_x^2 \sum_{k=0}^{N-1} h^2(k) \left(1 - \frac{k}{N + 1}\right) \quad (9.76)$$

Dalla (9.76) si possono ricavare numerose conclusioni. Innanzitutto risulta chiaro che la potenza di rumore in uscita è (come per il caso in virgola fissa) proporzionale ad N . Secondariamente, osserviamo che

$$\sigma_x^2 \sum_{k=0}^{N-1} h^2(k) \left(1 - \frac{k}{N + 1}\right) < \sigma_y^2$$

per cui il rapporto rumore-segnale in uscita per un ingresso con spettro piatto è limitato secondo la disequaglianza

$$\frac{\sigma_f^2}{\sigma_y^2} \leq (N + 1) \frac{2^{-2b}}{3} \quad (9.77)$$

Infine, ricordiamo dalla fig. 9.14(b) che si calcolavano i prodotti accumulando le somme in ordine di k crescente. È anche chiaro dalla (9.76) che i prodotti che si formano per primi sono quelli che risultano più fortemente pesati dal fattore $(1 - k/(N + 1))$. Ne deriva che la minima varianza del rumore di uscita si ottiene quando la risposta all'impulso soddisfa la relazione

$$|h(0)| < |h(1)| < \dots < |h(N - 1)|$$

Generalmente questa relazione non è soddisfatta. Tuttavia, se i prodotti vengono sommati in ordine di valore assoluto crescente della risposta al-

l'impulso, possiamo aspettarci di ottenere il più piccolo errore medio possibile in una realizzazione in virgola mobile. Ciò contrasta con il caso delle realizzazioni in virgola fissa, dove l'errore è indipendente dall'ordine delle addizioni. L'effettuare le moltiplicazioni e le somme parziali in ordine non sequenziale comporta generalmente ulteriore complessità di « software » o di « hardware »; tuttavia ne può valere la pena quando viene richiesta la massima precisione possibile. Esistono anche altri ordinamenti possibili che pure riducono il rumore di uscita (si veda il probl. 18 alla fine di questo capitolo).

9.5. EFFETTI DELLA LUNGHEZZA FINITA DEI REGISTRI NEI CALCOLI DELLA TRASFORMATA DI FOURIER DISCRETA

È importante capire gli effetti della lunghezza finita dei registri nei calcoli della trasformata di Fourier discreta, essendo questa ampiamente usata in pratica nel filtraggio numerico e nell'analisi spettrale. Come per il caso dei filtri numerici, tuttavia, un'analisi precisa di quegli effetti è difficile, e ci si limita pertanto ad analisi semplificate, spesso sufficienti per decidere la lunghezza di registro da usare nei calcoli della DFT. L'analisi che presenteremo è di tipo simile a quelle svolte nei precedenti paragrafi. In particolare, analizzeremo l'effetto dell'arrotondamento per mezzo di sorgenti di rumore additivo poste in tutti quei punti dell'algoritmo di calcolo dove si verifica l'arrotondamento. Faremo inoltre un certo numero di ipotesi per semplificare l'analisi. I risultati che se ne ricavano danno luogo a delle regole semplificate ma utili per gli accorgimenti da adottare nel calcolo della DFT rispetto agli effetti considerati. L'analisi che segue riguarda essenzialmente gli effetti dell'arrotondamento, ma, come per l'analisi dell'errore da arrotondamento nei filtri numerici, i risultati possono generalmente essere modificati per il caso del troncamento in complemento a due con l'aritmetica in virgola fissa, e per il troncamento in complemento a uno o in modulo e segno con l'aritmetica in virgola mobile.

Nel cap. 6 abbiamo visto che esistono diversi modi di effettuare il calcolo della DFT. In questo paragrafo considereremo innanzitutto gli errori nel calcolo diretto della DFT, e illustreremo poi gli effetti dell'arrotondamento in una classe particolare di algoritmi di FFT.

9.5.1 Analisi della quantizzazione nel calcolo della DFT

La trasformata di Fourier discreta è definita dall'espressione

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad k = 0, 1, \dots, N - 1 \quad (9.78)$$

dove $W_N = e^{-j(2\pi/N)k}$. Benché la (9.78) sia generalmente calcolata mediante uno di quegli algoritmi che vanno sotto il nome di *trasformata di Fourier veloce* (FFT), esistono diverse situazioni in cui risulta più conve-

niente accumulare direttamente i prodotti della (9.78); ciò, per esempio, si verifica quando si vuole conoscere la DFT solo per alcuni valori di k . L'analisi di questo procedimento di calcolo è piuttosto semplice e serve come introduzione allo studio degli effetti della quantizzazione nei calcoli della DFT.

Osserviamo che per un dato valore di k , la (9.78) è del tutto analoga alla somma di convoluzione

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k)$$

che è stata la base per lo sviluppo dell'analisi nel precedente paragrafo. In questo caso le quantità W_N^{kn} giocano il ruolo di risposta all'impulso, $X(k)$ gioca il ruolo dell'uscita, e $x(n)$ è l'ingresso. Si noti che tutte queste quantità sono generalmente complesse. Pertanto, per il calcolo diretto della DFT, si può usare un'analisi simile a quella del par. 9.4, tenendo presente che gli errori, in questo caso, sono delle sequenze complesse.

Per l'aritmetica in virgola fissa, un modo di effettuare il calcolo diretto di $X(k)$ è rappresentato dal grafo di flusso della fig. 9.15. Nella figura,

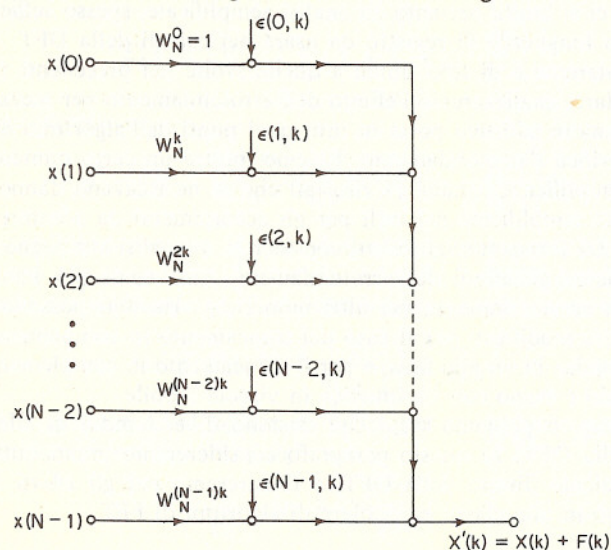


Fig. 9.15 Modello statistico per il rumore di arrotondamento in virgola fissa nel calcolo della DFT

$X'(k)$ rappresenta il risultato del calcolo della DFT con precisione finita, ed $F(k)$ rappresenta l'errore nel calcolo del k -mo valore. Le quantità complesse $\varepsilon(n, k)$ stanno a indicare gli errori dovuti all'arrotondamento dei prodotti $x(n)W_N^{kn}$. Si può vedere che questi errori complessi si sommano direttamente all'uscita, così che

$$F(k) = \sum_{n=0}^{N-1} \varepsilon(n, k) \quad (9.79)$$

Il prodotto $x(n)W_N^{kn}$ è

$$x(n)W_N^{kn} = \text{Re}[x(n)] \cos\left(\frac{2\pi}{N}kn\right) + \text{Im}[x(n)] \sin\left(\frac{2\pi}{N}kn\right) + j\left[\text{Im}[x(n)] \cos\left(\frac{2\pi}{N}kn\right) - \text{Re}[x(n)] \sin\left(\frac{2\pi}{N}kn\right)\right]$$

Con l'aritmetica in virgola fissa a precisione finita, i prodotti complessi arrotondati possono essere rappresentati come

$$\begin{aligned} Q[x(n)W_N^{kn}] &= \text{Re}[x(n)] \cos\left(\frac{2\pi}{N}kn\right) + \varepsilon_1(n, k) \\ &+ \text{Im}[x(n)] \sin\left(\frac{2\pi}{N}kn\right) + \varepsilon_2(n, k) \\ &+ j \text{Im}[x(n)] \cos\left(\frac{2\pi}{N}kn\right) + \varepsilon_3(n, k) \\ &- j \text{Re}[x(n)] \sin\left(\frac{2\pi}{N}kn\right) + \varepsilon_4(n, k) \end{aligned}$$

E cioè, ogni moltiplicazione reale contribuisce per un errore di arrotondamento⁵. Per calcolare la varianza dell'errore in $X'(k)$, dobbiamo fare una serie di ipotesi sugli errori. Più precisamente, assumeremo che gli errori dovuti ad ogni singola moltiplicazione reale abbiano le seguenti proprietà:

1. Gli errori sono variabili aleatorie uniformemente distribuite tra i valori $-\frac{1}{2} \cdot 2^{-b}$ e $\frac{1}{2} \cdot 2^{-b}$. Perciò, ogni sorgente di errore ha varianza $2^{-2b}/12$.
2. Gli errori sono scorrelati l'uno con l'altro.
3. Tutti gli errori sono scorrelati con l'ingresso e di conseguenza con l'uscita.

La media dell'errore dovuto all'arrotondamento di una moltiplicazione complessa è zero. Essendo il modulo al quadrato dell'errore complesso $\varepsilon(n, k)$ pari a

$$|\varepsilon(n, k)|^2 = [\varepsilon_1(n, k) + \varepsilon_2(n, k)]^2 + [\varepsilon_3(n, k) + \varepsilon_4(n, k)]^2$$

il valore medio di $|\varepsilon(n, k)|^2$ è

$$E[|\varepsilon(n, k)|^2] = 4 \cdot \frac{2^{-2b}}{12} = \frac{1}{3} \cdot 2^{-2b} \quad (9.80)$$

La media del modulo al quadrato dell'errore di uscita è

$$E[|F(k)|^2] = \sum_{n=0}^{N-1} E[|\varepsilon(n, k)|^2] = \frac{N}{3} 2^{-2b} \quad (9.81)$$

⁵ Si noti che abbiamo fatto l'ipotesi che i coefficienti W_N^{kn} siano rappresentati esattamente. L'effetto della quantizzazione di questi coefficienti sarà discusso nel par. 9.5.3.

Come nel caso della realizzazione in forma diretta di un filtro FIR, il rumore di uscita è proporzionale ad N^6 .

Come per la realizzazione in forma diretta e in virgola fissa di un filtro FIR, il calcolo diretto della DFT è soggetto a limitazioni nella dinamica della sequenza da trasformare. Dalla (9.78) vediamo che

$$|X(k)| \leq \sum_{n=0}^{N-1} |x(n)| < N$$

Affinché non si abbia saturazione occorre che sia $|X(k)| < 1$. Ciò è assicurato se

$$\sum_{n=0}^{N-1} |x(n)| < 1 \quad (9.82)$$

Pertanto, nel caso peggiore, potrà essere necessario dividere l'ingresso per N per evitare la saturazione. Per esempio, la sequenza $x(n) = 1$, per $0 \leq n \leq N-1$, ha la trasformata di Fourier discreta

$$X(k) = \begin{cases} N, & k = 0 \\ 0, & \text{altrove} \end{cases}$$

D'altra parte $|X(k)|$ può essere inferiore ad uno anche se la (9.82) non è soddisfatta, il che implica che la condizione (9.82) è sufficiente ma non necessaria affinché non si abbia saturazione.

Esistono diverse soluzioni al problema della dinamica. Si può dividere l'ingresso per N , aumentando in tal modo il rapporto rumore-segnale all'uscita. Si può usare uno schema di virgola mobile a blocchi dove si effettua la divisione per due tutte le volte che si ha una saturazione. Si può usare l'aritmetica in virgola mobile, nel qual caso la saturazione è virtualmente eliminata. Tutte queste possibilità saranno esaminate in dettaglio per una classe di algoritmi di FFT. In aggiunta, faremo anche qualche osservazione sull'errore introdotto dalla quantizzazione dei valori dei coefficienti.

9.5.2 Analisi degli effetti della quantizzazione negli algoritmi in virgola fissa [23]

Esistono numerosi algoritmi di FFT, e gli effetti della quantizzazione dipendono dallo specifico algoritmo usato. Quelli più comunemente adottati sono gli algoritmi basati sulla radice due, per i quali la dimensione della trasformata che viene calcolata è una potenza intera di due. Per la sua quasi totalità, la discussione che segue è imperniata sulla forma della decimazione nel tempo dell'algoritmo basato sulla radice due. I risultati, tuttavia, sono applicabili con poche modifiche alla forma della decimazio-

* Si noti che la (9.81) è una stima in eccesso in quanto alcune delle moltiplicazioni (per es. per W_N^0) possono essere fatte senza errore.

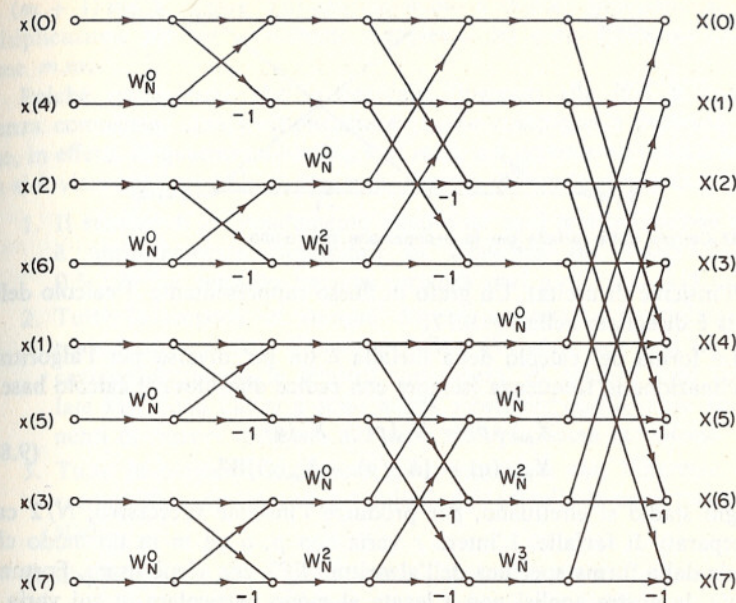


Fig. 9.16 Grafo di flusso per l'algoritmo di FFT di decimazione nel tempo

ne in frequenza. Inoltre, la maggior parte dei concetti utilizzati per l'analisi degli errori nel caso degli algoritmi basati sulla radice due sono utilizzabili anche per altri algoritmi.

Gli algoritmi di FFT servono a calcolare $X(k)$, la DFT di una sequenza finita $x(n)$, definita come nella (9.78). Nella fig. 9.16 è rappresentato il grafo di flusso di un algoritmo di decimazione nel tempo per $N = 8 = 2^3$ (una realizzazione di questa particolare forma dell'algoritmo è stata usata per il lavoro sperimentale al quale si è fatto cenno in precedenza). Esistono in questo schema alcuni aspetti chiave che, come ci ricordiamo dal cap. 6, sono comuni a tutti gli algoritmi standard basati sulla radice due. La DFT è calcolata in $v = \log_2 N$ stadi. Ad ogni stadio si forma un nuovo insieme ordinato di N numeri per mezzo di una combinazione lineare degli elementi dell'insieme precedente presi due alla volta. L'insieme v .mo contiene la DFT desiderata. Il calcolo numerico di base opera su di una coppia di numeri dell'insieme m .mo per produrre una coppia di numeri dell'insieme $(m+1)$.mo. Questo calcolo base, chiamato farfalla, è

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + W_N^r X_m(q) \\ X_{m+1}(q) &= X_m(p) - W_N^r X_m(q) \end{aligned} \quad (9.83)$$

Qui gli indici m ed $(m+1)$ si riferiscono agli insiemi m .mo ed $(m+1)$.mo rispettivamente, e p e q denotano la posizione dei numeri in ogni insieme (si noti che $m=0$ si riferisce all'insieme di ingresso ed $m=v$ si riferi-

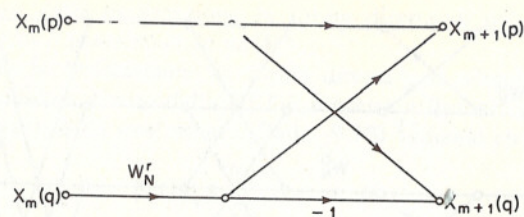


Fig. 9.17 Calcolo della farfalla per la decimazione nel tempo.

sce all'insieme di uscita). Un grafo di flusso rappresentante il calcolo della farfalla è disegnato nella fig. 9.17.

La forma del calcolo della farfalla è un po' diversa per l'algoritmo di decimazione in frequenza (sempre con radice due) dove il calcolo base è

$$\begin{aligned} X_{m+1}(p) &= X_m(p) + X_m(q) \\ X_{m+1}(q) &= [X_m(p) - X_m(q)]W_N^r \end{aligned} \quad (9.84)$$

Ad ogni stadio si effettuano, per produrre l'insieme successivo, $N/2$ calcoli separati di farfalle. L'intero r varia con p , q ed m in un modo che dipende dalla forma specifica dell'algoritmo FFT che viene usato. Fortunatamente, la nostra analisi non è legata al modo particolare in cui varia r . Anche la relazione specifica fra p , q ed m , che determina il modo in cui vanno sistemati gli indici nell'ambito dell' m .mo insieme, non è importante ai fini della nostra analisi. Nei due casi della decimazione nel tempo e della decimazione in frequenza i dettagli dell'analisi differiscono alquanto, a causa delle diverse forme della farfalla, ma i risultati fondamentali non cambiano in modo significativo. L'analisi che ora noi faremo sarà basata su di una farfalla di forma corrispondente alla (9.83) e cioè alla decimazione nel tempo.

Modelleremo il rumore di arrotondamento associando un generatore di rumore additivo ad ogni moltiplicazione in virgola fissa. Con questo modello la farfalla di fig. 9.17 viene sostituita, onde poter analizzare gli effetti del rumore di arrotondamento, con quella di fig. 9.18. Con la notazione $\epsilon(m, q)$ abbiamo esplicitamente voluto indicare il fatto che questa quantità rappresenta un errore complesso introdotto nel calcolo dell'insie-

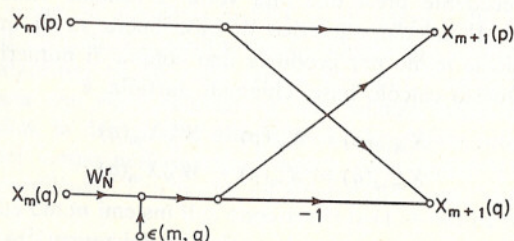


Fig. 9.18 Modello statistico per il rumore di arrotondamento in virgola fissa in un calcolo di farfalla per la decimazione nel tempo

me $(m+1)$.mo a partire dall'insieme m .mo e dovuto precisamente alla moltiplicazione per un coefficiente complesso del q .mo elemento dell'insieme m .mo.

Poiché assumeremo che in generale l'ingresso alla FFT è una sequenza complessa, ogni moltiplicazione risulta complessa e pertanto consiste, in effetti, di quattro moltiplicazioni reali, esattamente allo stesso modo che si è visto nel paragrafo precedente. Faremo inoltre le consuete ipotesi:

1. Il rumore di arrotondamento dovuto ad ogni moltiplicazione reale è uniformemente distribuito in ampiezza tra $-0.5 \cdot 2^{-b}$ e $0.5 \cdot 2^{-b}$ e pertanto ha una varianza pari a $\sigma_r^2 = (1/12) \cdot 2^{-2b}$.
2. Tutte le sorgenti di rumore dovute a ciascuna moltiplicazione reale sono scorrelate l'una con l'altra. Pertanto, per ogni moltiplicazione complessa le quattro componenti di rumore sono scorrelate l'una con l'altra e sono anche scorrelate con le altre componenti di rumore associate alle altre moltiplicazioni complesse.
3. Tutte le sorgenti di rumore sono scorrelate con l'ingresso e di conseguenza anche con i risultati dei calcoli di ogni insieme o stadio.

Poiché le quattro sequenze di rumore sono a media nulla e scorrelate tra loro, e hanno la stessa varianza, si ha, come nella (9.80),

$$E[|\epsilon(m, q)|^2] = \frac{1}{3} \cdot 2^{-2b} \quad (9.85)$$

Indicheremo questa varianza con σ_B^2 . Per calcolare il valore quadratico medio del rumore in ogni nodo di uscita, dobbiamo tener conto del contributo di ogni sorgente di rumore che si propaga fino a quel nodo. Dal grafo di flusso della fig. 9.16 possiamo trarre le seguenti osservazioni:

1. La funzione di trasferimento da ogni nodo del grafo di flusso a ogni altro nodo cui è connesso, è la moltiplicazione per una costante complessa di ampiezza uno (in quanto il coefficiente di trasmissione in ogni ramo è o uno o una potenza intera di W_N).
2. Ogni nodo di uscita è connesso nel grafo di flusso a $7 = (N-1)$ farfalle. La fig. 9.19(a) mostra, per esempio, il grafo di flusso una volta eliminate tutte le farfalle che non sono connesse a $X(0)$, e la fig. 9.19(b) mostra la stessa cosa rispetto a $X(2)$.

Le osservazioni fatte sopra si possono generalizzare al caso in cui N è una potenza arbitraria di 2.

Come conseguenza della prima osservazione si ha che il valore quadratico medio dell'ampiezza della componente del rumore di uscita dovuta ad ogni sorgente elementare di rumore è lo stesso ed è uguale a σ_B^2 . Il rumore totale ad ogni nodo di uscita è dato dai contributi di rumore che si propagano fino a quel nodo. Avendo assunto che tutte le sorgenti di rumore sono scorrelate, il valore quadratico medio dell'ampiezza del rumore di

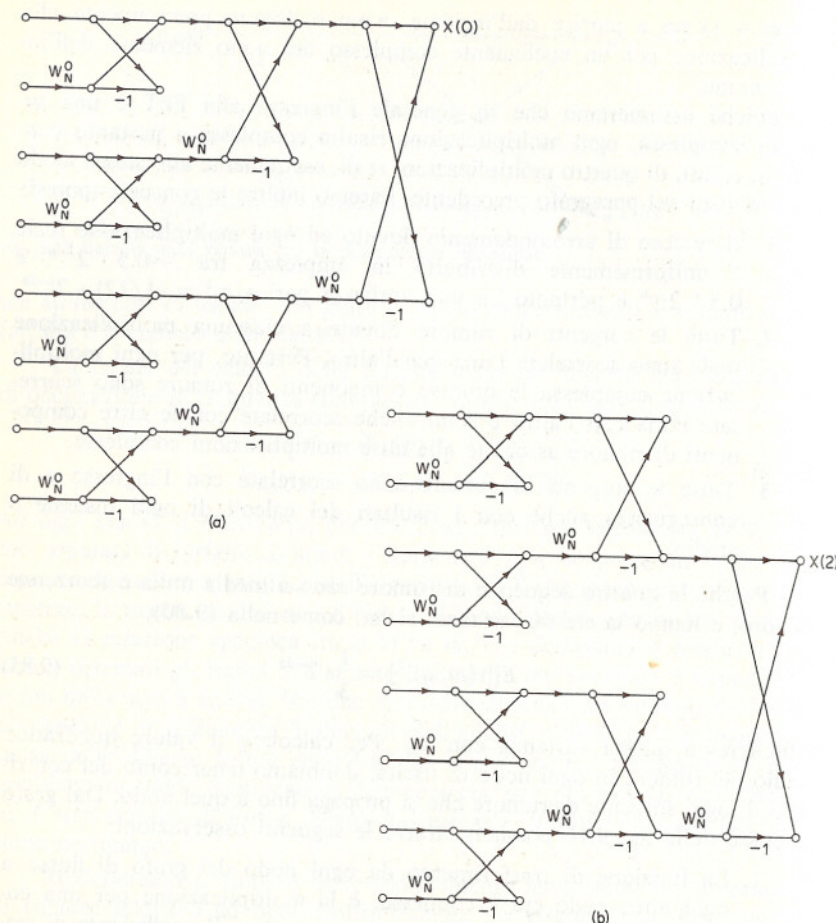


Fig. 9.19 (a) Farfalle che influiscono su $X(0)$; (b) farfalle che influiscono su $X(2)$.

uscita è uguale a σ_B^2 moltiplicato il numero delle sorgenti di rumore che si propagano fino a quel nodo. Per ogni farfalla si introduce al più una sorgente di rumore complesso; di conseguenza, in base alla osservazione no.2 fatta sopra, si propagano fino ad ogni nodo di uscita al più $(N-1)$ sorgenti di rumore. In realtà, non tutte le farfalle generano rumore di arrotondamento, poiché alcune (per esempio tutte quelle nel primo e nel secondo stadio) comportano solo moltiplicazioni per uno. Se assumiamo, tuttavia, che l'arrotondamento si effettui in ogni farfalla, potremo considerare il risultato come un limite superiore per il rumore di uscita. Pertanto, con questa ipotesi, il valore quadratico medio del rumore di uscita $F(k)$, in corrispondenza del k -mo valore della DFT, è dato da

$$E[|F(k)|^2] = (N-1)\sigma_B^2 \quad (9.86)$$

che, per N grande, approssimeremo con

$$E[|F(k)|^2] \cong N\sigma_B^2 \quad (9.87)$$

Secondo questo risultato, il valore quadratico medio del rumore di uscita è proporzionale ad N , il numero dei punti trasformati. Raddoppiare N , o aggiungere un altro stadio nella FFT, ha l'effetto di raddoppiare il valore quadratico medio del rumore di uscita. Nel probl. 9.19 considereremo il modo in cui si modifica questo risultato quando non si inseriscono le sorgenti di rumore per quelle farfalle che comportano soltanto la moltiplicazione per uno o per j .

Nel realizzare un algoritmo di FFT in virgola fissa occorre prendere misure nei confronti della saturazione. Dalla (9.83) segue che (v. probl. 2 del cap. 6)

$$\max(|X_m(p)|, |X_m(q)|) \leq \max(|X_{m+1}(p)|, |X_{m+1}(q)|) \quad (9.88)$$

ed anche che

$$\max(|X_{m+1}(p)|, |X_{m+1}(q)|) \leq 2 \max(|X_m(p)|, |X_m(q)|) \quad (9.89)$$

La (9.88) implica che il modulo massimo è non decrescente da stadio a stadio, così che, se il valore assoluto dell'uscita della FFT è minore di uno, allora il valore assoluto dei valori in ogni insieme o stadio deve essere minore di uno⁷; vale a dire che non si avrà saturazione in nessuno degli insiemi.

Per esprimere questa condizione con una limitazione sulla sequenza di ingresso, ricorderemo dal paragrafo precedente che la condizione

$$|x(n)| < \frac{1}{N}, \quad 0 \leq n \leq N-1 \quad (9.90)$$

è sia necessaria che sufficiente per garantire che

$$|X(k)| < 1, \quad 0 \leq k \leq N-1$$

Pertanto la (9.90) è sufficiente per garantire che non si abbia saturazione in ciascuno stadio dell'algoritmo.

Per ottenere un'espressione esplicita per il rapporto rumore-segnale all'uscita dell'algoritmo di FFT, tenendo conto al tempo stesso del controllo di scala richiesto, si consideri un ingresso per il quale valori successivi della sequenza siano scorrelati, e cioè un segnale di ingresso bianco. Si assuma inoltre che le parti reale ed immaginaria della sequenza di ingresso siano scorrelate e che ciascuna di esse abbia una densità di probabilità uniforme tra $-1/(\sqrt{2}N)$ e $+1/(\sqrt{2}N)$. Notare che questo segnale soddisfa la

⁷ In realtà si dovrebbe considerare la saturazione in termini di parte reale e parte immaginaria dei dati, piuttosto che in termini di modulo. Tuttavia, $|x| < 1$ implica $|\operatorname{Re}(x)| < 1$ e $|\operatorname{Im}(x)| < 1$, mentre è stato verificato che controllando la scala sulla base delle parti Re e Im si otterrebbe solo un leggero aumento del livello accettabile di segnale.

(9.90). In tal caso il valore quadratico medio della sequenza complessa di ingresso è

$$E[|x(n)|^2] = \frac{1}{3N^2} = \sigma_x^2 \quad (9.91)$$

La DFT della sequenza di ingresso è

$$X(k) = \sum_{n=0}^{N-1} x(n)W^{kn}$$

da cui si può dimostrare che, sotto le ipotesi di prima sull'ingresso,

$$\begin{aligned} E[|X(k)|^2] &= \sum_{k=0}^{N-1} E[|x(n)|^2] |W^{kn}|^2 \\ &= N\sigma_x^2 = \frac{1}{3N} \end{aligned} \quad (9.92)$$

Mettendo insieme la (9.87) e la (9.92) otteniamo

$$\frac{E[|F(k)|^2]}{E[|X(k)|^2]} = 3N^2\sigma_B^2 = N^2 2^{-2b} \quad (9.93)$$

Pertanto, stando a questo risultato, il rapporto rumore-segnale è proporzionale a N^2 . Essendo σ_B^2 proporzionale a 2^{-2b} , la (9.93) implica che il rapporto rumore-segnale cresce come N^2 oppure di un bit per stadio: vale a dire che, se N raddoppia, il che corrisponde ad aggiungere uno stadio alla FFT, allora, per mantenere lo stesso rapporto rumore-segnale, occorre aggiungere un bit alla lunghezza dei registri. L'ipotesi di un segnale di ingresso bianco non è critica a questo riguardo. Per una quantità di altri ingressi, infatti, il rapporto rumore-segnale è ancora proporzionale ad N^2 e cambia solo la costante di proporzionalità.

La (9.89) suggerisce un procedimento alternativo per la correzione di scala. Poiché il valore assoluto massimo aumenta da stadio a stadio di non più di un fattore due, si può evitare la saturazione imponendo che sia $|x(n)| < 1$ e inserendo un fattore di attenuazione di $1/2$ all'inizio di ogni stadio. In tal caso l'uscita non sarà la DFT definita dalla (9.78), ma $1/N$ volte questa DFT. Se da un lato il valore quadratico medio del segnale di uscita sarà $1/N^2$ volte più piccolo, dall'altro l'ampiezza dell'ingresso potrà essere N volte maggiore senza causare saturazione. Pertanto l'ampiezza di uscita massima che si può raggiungere (per un ingresso bianco) è la stessa di prima. Tuttavia, il livello del rumore in uscita sarà molto minore che nella (9.87), in quanto il rumore introdotto nei primi stadi della FFT sarà attenuato dalla correzione di scala che ha luogo negli stadi successivi. In particolare, con la correzione di scala di $1/2$ introdotta all'ingresso di ogni farfalla, lo schema della fig. 9.18 viene modificato in quello della fig. 9.20, dove, si noti, ci sono ora due sorgenti di rumore associate ad ogni farfalla. Come in precedenza, assumiamo che le parti reale ed immaginaria di queste sorgenti di rumore siano scorrelate tra loro

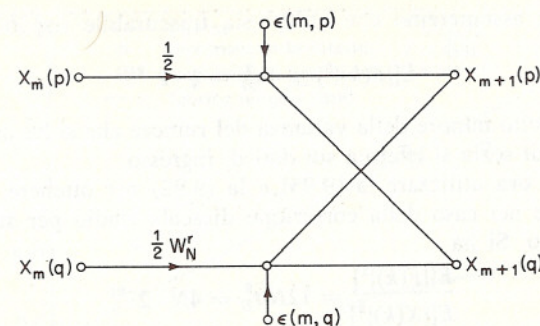


Fig. 9.20 Schema di farfalla contenente i moltiplicatori per la correzione di scala con associato rumore di arrotondamento in virgola fissa.

e con le altre sorgenti di rumore, e che le parti reale ed immaginaria siano uniformemente distribuite tra $-1/2 \cdot 2^{-b}$ e $+1/2 \cdot 2^{-b}$. Pertanto, come prima,

$$E[|\epsilon(m, q)|^2] = \sigma_B^2 = \frac{1}{3} \cdot 2^{-2b} = E[|\epsilon(m, p)|^2]$$

Essendo le sorgenti di rumore tutte scorrelate, il valore quadratico medio del modulo del rumore in ogni nodo di uscita è ancora la somma dei contributi di ciascuna sorgente di rumore nel grafo di flusso. Tuttavia, al contrario del caso precedente, l'attenuazione cui va incontro ogni sorgente di rumore attraverso il grafo di flusso dipende dal particolare stadio di origine. Una sorgente di rumore che ha origine all'uscita dello stadio m si propaga fino all'uscita moltiplicandosi per una costante complessa di modulo $(1/2)^{v-m-1}$. Dall'esame della fig. 9.16 si deduce che, per il caso $N = 8$, ogni nodo di uscita si connette a

- 1 farfalla con origine allo stadio $(v-1)$.mo
- 2 farfalle con origine allo stadio $(v-2)$.mo
- 4 allo stadio $(v-3)$.mo, ecc.

Per il caso generale con $N = 2^v$, ogni nodo di uscita si connette a $2^{(v-m-1)}$ farfalle e perciò a $2^{(v-m)}$ sorgenti di rumore con origine all'uscita dello stadio m .mo. Pertanto, in ogni nodo di uscita il valore quadratico medio del modulo del rumore è

$$\begin{aligned} E[|F(k)|^2] &= \sigma_B^2 \sum_{m=0}^{v-1} 2^{(v-m)} \left(\frac{1}{2}\right)^{(2v-2m-2)} \\ &= \sigma_B^2 \sum_{m=0}^{v-1} \left(\frac{1}{2}\right)^{(v-m-2)} \\ &= \sigma_B^2 \cdot 2 \sum_{k=0}^{v-1} \left(\frac{1}{2}\right)^k \\ &= 2\sigma_B^2 \frac{1 - (\frac{1}{2})^v}{1 - \frac{1}{2}} = 4\sigma_B^2 (1 - (\frac{1}{2})^v) \end{aligned} \quad (9.94)$$

Per N grande assumeremo che $(1/2)^v$ sia trascurabile rispetto a 1, così che

$$E[|F(k)|^2] \cong 4\sigma_B^2 = \frac{4}{3} \cdot 2^{-2b} \quad (9.95)$$

ed è quindi molto minore della varianza del rumore che si ha quando tutta la correzione di scala si effettua sui dati di ingresso.

Possiamo ora utilizzare la (9.95) e la (9.92) per ottenere il rapporto rumore-segnale nel caso della correzione di scala stadio per stadio e con ingresso bianco. Si ha

$$\frac{E[|F(k)|^2]}{E[|X(k)|^2]} = 12N\sigma_B^2 = 4N \cdot 2^{-2b} \quad (9.96)$$

un risultato proporzionale ad N invece che ad N^2 . Un'interpretazione della (9.96) è che il rapporto rumore-segnale in uscita cresce come N , o di metà bit per stadio, come dimostrato per la prima volta da Welch [23]. È importante osservare anche qui che l'ipotesi di segnale bianco non è essenziale nell'analisi. Il risultato fondamentale dell'incremento di mezzo bit per stadio vale per un'ampia classe di segnali, venendo a cambiare nella (9.96) solo la costante moltiplicativa.

Va ancora osservato che la causa principale dell'aumentare con N del rapporto rumore-segnale è la diminuzione del livello di segnale (imposta dai vincoli sulla saturazione) quando si passa da uno stadio all'altro. Secondo la (9.95), pochissimo rumore (solo un bit o due) è presente nello stadio finale. La maggior parte del rumore è stato eliminato grazie alle moltiplicazioni per i fattori di scala.

La precedente discussione si è basata sull'ipotesi di calcoli svolti rigidamente in virgola fissa, comportanti, cioè, solo attenuazioni prestabilite e non correzioni di scala basate su test di saturazione. Chiaramente, se i circuiti o i mezzi di programmazione disponibili sono tali da richiedere una rigida attuazione dei calcoli in virgola fissa, allora si dovrebbero possibilmente incorporare in ogni stadio i fattori di attenuazione pari a $1/2$, anziché usare una forte attenuazione per il solo stadio di ingresso.

Un terzo modo per evitare la saturazione è basato sull'uso della virgola mobile a blocchi. In questo procedimento l'insieme originario dei campioni (corrispondente al primo stadio) è normalizzato all'estrema sinistra del registro, con la restrizione che sia $|x(n)| < 1$; il calcolo procede col sistema a virgola fissa, ma dopo ogni addizione c'è un test di saturazione. Se questa viene rivelata, l'intero insieme di valori di quello stadio viene diviso per due e il calcolo prosegue. Il numero delle divisioni necessarie viene contato in modo da determinare il fattore di scala o esponente per l'intero insieme finale di valori. Il rapporto rumore-segnale dipende fortemente da quante saturazioni si verificano e da dove si verificano. Le posizioni e l'ordine con cui si verificano le saturazioni dipendono dal segnale che viene trasformato, e pertanto, al fine di analizzare il rapporto rumore-segnale in una realizzazione di FFT in virgola mobile a blocchi, è necessaria conoscere le proprietà del segnale di ingresso.

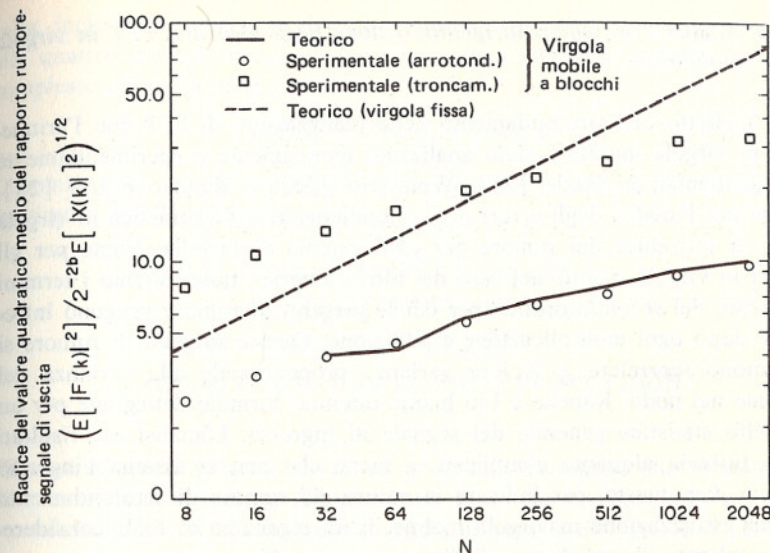


Fig. 9.21 Rapporto rumore-segnale di uscita, sperimentale, per una realizzazione di FFT in virgola mobile a blocchi.

Come si è già visto, è possibile trovare sia un ingresso che non richiede affatto correzioni di scala, sia un ingresso che richiede la divisione per N per evitare la saturazione. Ci si può aspettare che il caso dell'ingresso costituito da un rumore bianco fornisca un esempio situato in qualche modo a metà strada fra i due estremi, e, cioè, tale per cui la correzione di scala non sia generalmente necessaria ad ogni stadio. Questo problema è stato analizzato teoricamente [19], ma l'analisi è alquanto complicata e ci limiteremo pertanto a presentare dei risultati sperimentali.

La fig. 9.21 illustra la dipendenza da N del rapporto rumore-segnale in uscita. Questa figura mostra i valori del rapporto rumore-segnale misurati sperimentalmente per segnali di ingresso bianchi e per trasformate in virgola mobile a blocchi con arrotondamento [26]. Per confronto è anche mostrata la curva teorica rappresentante il rapporto-rumore-segnale in virgola fissa corrispondente alla (9.96). Si può vedere che, per questo tipo di ingresso, la virgola mobile a blocchi consente qualche vantaggio sulla virgola fissa, specialmente per le trasformate più grosse. Per $N = 2048$ il rapporto rumore-segnale per la virgola mobile a blocchi è circa $1/8$ rispetto al caso della virgola fissa, corrispondente a un miglioramento di tre bit.

Per vedere come cambiano i risultati quando in virgola mobile a blocchi si usa il troncamento al posto dell'arrotondamento, si è fatto ricorso a un'indagine sperimentale [26]. I risultati di questo esperimento sono anch'essi mostrati nella fig. 9.21. I rapporti rumore-segnale sono generalmente un poco peggiori che per il caso dell'arrotondamento. Il ritmo di crescita del rapporto rumore-segnale con N sembra essere pressoché lo stesso che per l'arrotondamento.

9.5.3 Analisi degli effetti di quantizzazione negli algoritmi FFT in virgola mobile

L'effetto dell'arrotondamento nelle realizzazioni di FFT con l'aritmetica in virgola mobile è stato analizzato teoricamente e sperimentalmente da Gentleman e Sande [24], Weinstein [26], e Kaneko e Liu [25]. Come per l'analisi degli errori di arrotondamento nell'aritmetica in virgola fissa, si introduce del rumore per ogni calcolo di farfalla. Come per gli errori in virgola mobile nel caso dei filtri numerici, trascureremo i termini di errore del secondo ordine, per cui le sorgenti di rumore vengono introdotte dopo ogni moltiplicazione e addizione. Queste sorgenti di rumore si assumono scorrelate, e la loro varianza proporzionale alla varianza del segnale nel nodo. Kaneko e Liu hanno ottenuto formule dettagliate per un modello statistico generale del segnale di ingresso. L'analisi e i risultati sono tuttavia alquanto complicati, a meno che non si assuma l'ingresso bianco. Per questo, per indicare la natura del rumore di arrotondamento in una realizzazione in virgola mobile di un algoritmo di FFT, considereremo soltanto il caso di segnali di ingresso bianchi.

La fig. 9.22 rappresenta la metà superiore di un tipico calcolo di far-

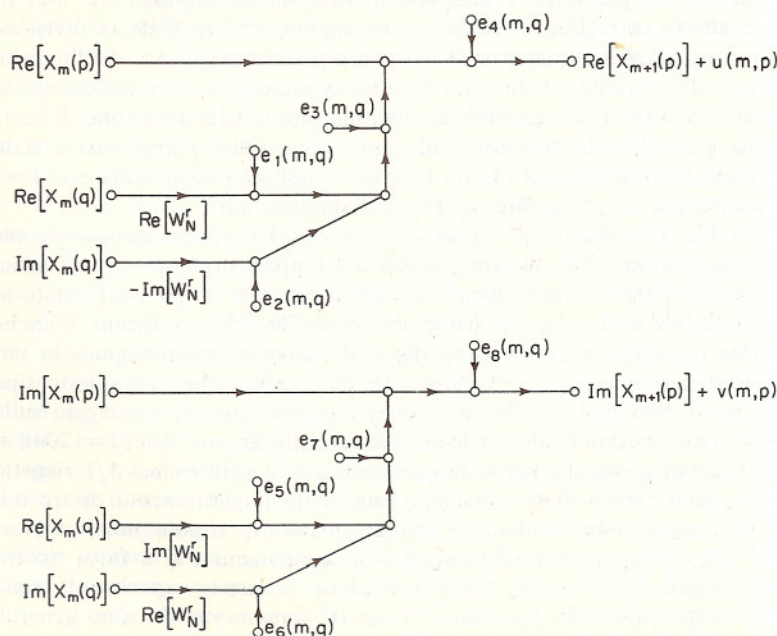


Fig. 9.22 Calcolo di farfalla con rumore per l'aritmetica in virgola mobile.

falla, includente le sorgenti di rumore reale $e_1(m, q), \dots, e_8(m, q)$ dovute alle quattro moltiplicazioni reali e alle quattro addizioni reali. Il rumore complesso all'uscita della farfalla è

$$s(m, p) = u(m, p) + jv(m, p)$$

Coerentemente con la nostra precedente discussione dei modelli di calcoli in virgola mobile, possiamo scrivere

$$\begin{aligned} e_1(m, q) &= \varepsilon_1(m, q) \operatorname{Re} [W_N^r] \operatorname{Re} [X_m(q)] \\ e_2(m, q) &= -\varepsilon_2(m, q) \operatorname{Im} [W_N^r] \operatorname{Im} [X_m(q)] \\ e_3(m, q) &= \varepsilon_3(m, q) \operatorname{Re} [W_N^r X_m(q)] \\ e_4(m, q) &= \varepsilon_4(m, q) \{ \operatorname{Re} [X_m(p)] + \operatorname{Re} [W_N^r X_m(q)] \} \end{aligned} \quad (9.97)$$

Espressioni simili valgono per e_5, e_6, e_7 ed e_8 . Assumiamo che le sorgenti di rumore bianco $\varepsilon_i(m, q)$ abbiano tutte la stessa varianza (σ_ε^2) e siano scorrelate l'una con l'altra. Inoltre, poiché il segnale di ingresso è bianco, con varianze uguali per le parti reale ed immaginaria, saranno bianche e con uguale varianza anche le parti reale ed immaginaria dei segnali in ogni stadio della FFT. Possiamo allora scrivere

$$E[(\operatorname{Re} [X_m(q)])^2] = E[(\operatorname{Im} [X_m(q)])^2] = \frac{1}{2} E[|X_m(q)|^2] \quad (9.98)$$

Dalle (9.97) e (9.98) segue che

$$\begin{aligned} \sigma_{e_1}^2 + \sigma_{e_2}^2 &= \sigma_{e_5}^2 + \sigma_{e_6}^2 = \sigma_{e_3}^2 = \sigma_{e_7}^2 = \frac{1}{2} \sigma_\varepsilon^2 E[|X_m(q)|^2] \\ \sigma_{e_4}^2 &= \sigma_{e_8}^2 = \sigma_\varepsilon^2 E[|X_m(q)|^2] \end{aligned}$$

I valori quadratici medi di $u(m, p)$ e $v(m, p)$ sono quindi

$$E[(v(m, p))^2] = E[(u(m, p))^2] = 2\sigma_\varepsilon^2 E[|X_m(q)|^2]$$

per cui il valore quadratico medio del modulo della sorgente di rumore in uscita $s(m, p)$ è

$$E[|s(m, p)|^2] = 4\sigma_\varepsilon^2 E[|X_m(q)|^2]$$

Pertanto la varianza del rumore generato nel calcolare l'insieme $(m+1)$.mo di valori è $4\sigma_\varepsilon^2$ volte il valore quadratico medio del segnale all'uscita dello stadio precedente. Se l'ingresso ($m=0$) è bianco e con valore quadratico medio $E[|x(n)|^2]$, allora il rumore generato nel calcolare l'insieme $(m+1)$.mo di valori è $2^m E[|x(n)|^2] (4\sigma_\varepsilon^2)$. Come prima, ogni nodo di uscita si connette a $2^{(v-m-1)}$ farfalle che appartengono allo stadio $(m+1)$.mo e si connette quindi a $2^{(v-m-1)}$ sorgenti di rumore associate all'uscita dello stadio $(m+1)$.mo. Ciascuna di queste sorgenti di rumore si propaga fino all'uscita venendo moltiplicata per una costante complessa

di modulo unitario. Pertanto il valore quadratico medio del rumore di uscita è

$$\begin{aligned} E[|F(k)|^2] &= \sum_{m=0}^{v-1} 2^{(v-m-1)} 2^m E[|x(n)|^2] (4\sigma_e^2) \\ &= \sum_{m=0}^{v-1} \frac{N}{2} E[|x(n)|^2] (4\sigma_e^2) \\ &= 2vN\sigma_e^2 E[|x(n)|^2] \end{aligned} \quad (9.99)$$

Il valore quadratico medio del segnale di uscita è

$$E[|X(k)|^2] = E[|x(n)|^2]N$$

e quindi il rapporto rumore-segnale in uscita è

$$\frac{E[|F(k)|^2]}{E[|X(k)|^2]} = 2v\sigma_e^2 \quad (9.100)$$

Dalla (9.100) osserviamo che il rapporto rumore-segnale è proporzionale a v , mentre nel caso della virgola fissa era proporzionale a $N = 2^v$. Poiché σ_e^2 è proporzionale a 2^{-2b} , quadruplicando v , (cioè elevando N alla quarta potenza) si ha un aumento di un bit nel rapporto rumore-segnale. Pertanto, proprio come ci si aspetta, l'aumento del rapporto rumore-segnale come funzione di N nel caso della virgola mobile è alquanto più debole che nel caso della virgola fissa.

Nella analisi che ha portato alla (9.100), abbiamo trascurato il fatto che le moltiplicazioni per uno possono effettuarsi senza rumore. Per un determinato algoritmo di FFT basato sulla radice 2, come per esempio l'algoritmo di decimazione nel tempo illustrato nella fig. 9.16, si può tener conto della riduzione di varianza dovuta all'essere, per certi r , $W_N^r = 1$ e j , ottenendo una stima leggermente inferiore del rapporto rumore-segnale in uscita. Tuttavia, per N sufficientemente grande, questa analisi più corretta del rumore dà luogo a previsioni sul rumore di uscita solo di poco migliori dell'analisi semplificata svolta sopra.

I risultati discussi sopra sono in eccellente accordo con le verifiche fatte da Weinstein [26], come dimostra la fig. 9.23. Per ottenere tale accordo, tuttavia, è stato necessario rendere casuale l'arrotondamento, e cioè arrotondare a caso in su e in giù quando il valore della mantissa era esattamente $0.5 \cdot 2^{-b}$. La curva teorica modificata della figura è stata ottenuta utilizzando la riduzione dei valori della varianza del rumore dovuta all'essere, per certi r , $W_N^r = 1$ e $W_N^j = j$. La figura mostra anche i risultati sperimentali per l'arrotondamento non casuale. Tali risultati sono stati interpolati con una curva della forma av^2 , ma una simile relazione quadratica non è stata stabilita teoricamente.

Tutto quello che abbiamo visto finora, risultati sperimentali compresi, vale per il caso di un segnale bianco. È stata tuttavia svolta qualche indagine sperimentale per verificare se le previsioni effettuate valgono anche

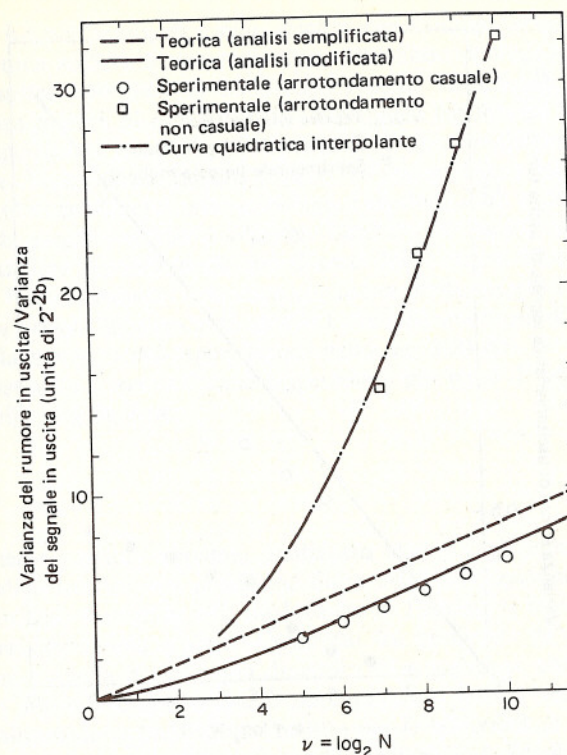


Fig. 9.23 Rapporto rumore-segnale sperimentale e teorico per calcoli di FFT in virgola mobile

senza questo vincolo. In particolare, è stato misurato il rumore introdotto nel calcolo della FFT di segnali sinusoidali di diverse frequenze, per $v = 8, 9, 10$ e 11 . I risultati, mediati sulle frequenze di ingresso usate, sono stati entro il 15% rispetto a quelli previsti dalla (9.100). Tali esperimenti sono stati effettuati utilizzando l'arrotondamento « casuale ».

9.5.4 Effetti della quantizzazione dei coefficienti nella FFT

Come per i filtri numerici, anche la realizzazione di un algoritmo di FFT richiede l'uso di coefficienti quantizzati. Benché la quantizzazione dei coefficienti sia intrinsecamente un fatto non statistico, Weinstein [26] ha ottenuto alcuni risultati utili per mezzo di un modello statistico alquanto approssimativo. In tale modello ogni coefficiente viene sostituito dal suo vero valore più una sequenza di rumore bianco, il che equivale ad aggiungere del « jitter » ad ogni coefficiente. Sebbene a rigore l'effetto dell'errore dovuto alla quantizzazione sia diverso da quello dovuto al jitter, è ragio-

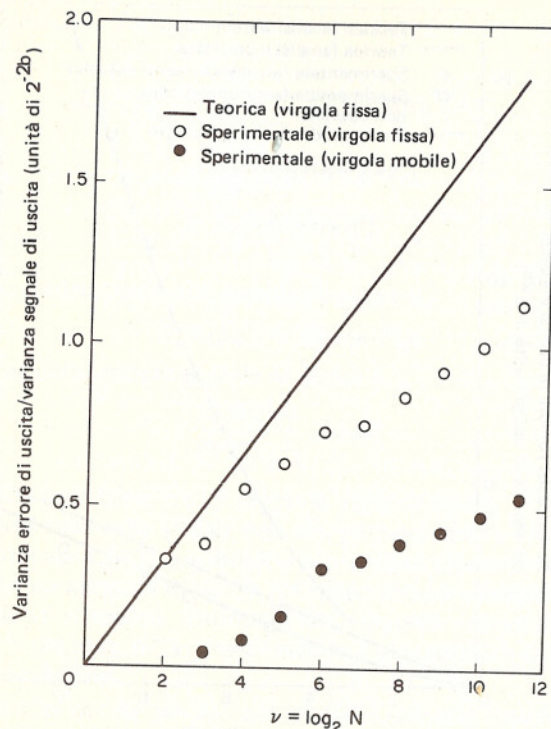


Fig. 9.24 Errori dovuti alla quantizzazione dei coefficienti nei calcoli della FFT.

nevole attendersi che, in prima approssimazione, l'ampiezza degli errori nei due casi sia confrontabile. Il risultato teorico ottenuto da Weinstein è che il rapporto tra i valori quadratici medi dell'errore e del segnale in uscita è $(\nu/6) \cdot 2^{-2b}$.

Sebbene questo risultato non permetta di prevedere con grande precisione l'errore dovuto alla quantizzazione dei coefficienti in un algoritmo di FFT, esso serve tuttavia come stima approssimativa dell'errore. Il risultato sostanziale, dimostrato anche sperimentalmente, è che il rapporto rumore-segnale cresce in modo molto blando con N , essendo proporzionale a $\nu = \log_2 N$, così che un raddoppio di N produce soltanto un leggero aumento nel rapporto rumore-segnale.

I risultati sperimentali sono illustrati nella fig. 9.24; la quantità rappresentata in ordinata è 2^{2b} volte il rapporto tra i valori quadratici medi dell'errore e del segnale di uscita. La curva teorica è disegnata a tratto pieno, mentre i circoletti rappresentano le misure sperimentali del rapporto rumore-segnale in uscita nel caso della virgola fissa. I risultati sperimentali si situano generalmente sotto la curva teorica. Nessun risultato sperimentale differisce per più di un fattore due dal risultato teorico, e siccome un fattore due nel rapporto rumore-segnale corrisponde a una differenza di

solo mezzo bit nell'errore di uscita, sembra che l'analisi in questione fornisca una stima sufficientemente accurata dell'effetto dell'errore nei coefficienti. I risultati sperimentali sembrano crescere linearmente con ν , ma con pendenza minore di quella fornita dall'analisi teorica.

Gli esperimenti citati sopra sono relativi al caso dell'aritmetica in virgola fissa. Tuttavia, poiché una FFT in virgola mobile a blocchi fa uso generalmente di coefficienti in virgola fissa, i risultati sono validi anche per il caso della virgola mobile a blocchi. Con alcune leggere modifiche, è possibile ottenere risultati simili per il caso della virgola mobile. Eccetto che per un fattore costante, i risultati per il caso della virgola mobile e per quello della virgola fissa sono gli stessi. I risultati sperimentali per il caso della virgola mobile sono rappresentati nella fig. 9.24 dai cerchietti pieni e si può osservare come essi siano leggermente più bassi di quelli ottenuti nel caso della virgola fissa.

SOMMARIO

In questo capitolo abbiamo analizzato alcuni degli effetti derivanti dall'uso dell'aritmetica con precisione finita nella realizzazione degli algoritmi di filtraggio numerico e di trasformate di Fourier. Il tema dominante della nostra discussione è stato il conflitto fra l'esigenza di una quantizzazione fine e quella di un'ampia dinamica di segnale per una fissata lunghezza dei registri. Quest'ultima è un importante fattore economico per tutte le realizzazioni « hardware », mentre per le realizzazioni « software » essa è generalmente imposta dalle caratteristiche del calcolatore di uso generale disponibile. È pertanto molto utile capire gli effetti della quantizzazione in tutte le applicazioni di elaborazione numerica dei segnali.

Le nostre considerazioni sugli effetti della lunghezza finita dei registri sono cominciate con la discussione dei vari tipi di rappresentazione dei numeri solitamente usati nelle realizzazioni degli algoritmi di elaborazione numerica dei segnali. Si sono quindi considerati in particolare gli algoritmi di filtraggio numerico e quelli di FFT. Il nostro obiettivo principale lungo tutto lo svolgimento del capitolo è stato quello di mettere in evidenza alcuni dei problemi risultanti dall'uso dell'aritmetica con precisione finita e di illustrare un tipo di analisi che può essere applicata con successo in problemi specifici di quantizzazione in algoritmi di elaborazione numerica dei segnali. Dagli esempi illustrativi che compongono questo capitolo emergono un certo numero di principi e suggerimenti generali. Abbiamo dimostrato, per esempio, l'esistenza dei cicli limite per ingresso zero nelle realizzazioni ricorsive di filtri numerici IIR e abbiamo dato alcune semplici formule che consentono di prevedere la dimensione di eventuali cicli limite per ingresso zero. Nel caso di ingressi non semplici, si è visto come una analisi statistica può consentire di stimare utilmente gli effetti della quantizzazione in termini di rapporti rumore-segnale sia per i filtri numerici che per gli algoritmi di FFT. Abbiamo inoltre considerato tanto le realizza-

zioni in virgola fissa che quelle in virgola mobile, e ciò sia per i filtri IIR e FIR che per gli algoritmi di FFT. Un'osservazione fondamentale è che, sia per le realizzazioni in virgola fissa che per quelle in virgola mobile, gli effetti della quantizzazione dipendono in buona misura dalla forma o struttura scelta per realizzare un particolare algoritmo di elaborazione di segnali. Si è visto inoltre come gli effetti della quantizzazione nelle realizzazioni in virgola mobile dipendono dalle proprietà dei dati di ingresso, mentre ciò non è generalmente vero per le realizzazioni in virgola fissa.

Al di là dei risultati specifici ottenuti, molti dei quali di vasta applicabilità, l'intero capitolo serve ad illustrare un tipo di analisi che si può applicare allo studio degli effetti della quantizzazione in una gran quantità di algoritmi di elaborazione numerica dei segnali. L'uso di metodi statistici per problemi dove i processi sono sconosciuti o troppo complessi per una rappresentazione deterministica, è un approccio ampiamente affermato in numerosi campi, incluso quello dell'elaborazione numerica dei segnali. In definitiva, gli esempi trattati in questo capitolo indicano il tipo di ipotesi e di approssimazioni che si fanno comunemente nello studio degli effetti della quantizzazione. Notiamo infine che diversi metodi di analisi di e con rumore sono anche illustrati in articoli di rassegna di Liu [27] e Oppenheim e Weinstein [28].

BIBLIOGRAFIA

1. I. Flores, *The Logic of Computer Arithmetic*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.
2. W. R. Bennett, "Spectra of Quantized Signals," *Bell System Tech. J.*, Vol. 27, 1948, pp. 446-472.
3. B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," *IRE Trans. Circuit Theory*, Vol. CT-3, Dec. 1956, pp. 266-276.
4. B. Widrow, "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems," *AIEE Trans. (Appl. Indust.)*, Vol. 81, Jan. 1961, pp. 555-568.
5. L. Jackson, "An Analysis of Limit Cycles Due to Multiplication Rounding in Recursive Digital Filters," in *Proc. 7th Allerton Conf. Circuit System Theory*, 1969, pp. 69-78.
6. L. B. Jackson, "An Analysis of Roundoff Noise in Digital Filters," Sc.D. Dissertation, Department of Electrical Engineering, Stevens Institute of Technology, 1969.
7. A. R. Parker and S. F. Hess, "Limit-Cycle Oscillations in Digital Filters," *IEEE Trans. Circuit Theory*, Vol. CT-8, Nov. 1971, pp. 687-697.
8. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," in *Proc. 7th Annual Allerton Conf. Circuit System Theory*, 1968, pp. 63-68.
9. I. W. Sandberg and J. F. Kaiser, "A Bound on Limit Cycles in Fixed-Point Implementations of Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 2, June 1972, pp. 110-112.
10. C. Y. Kao, "An Analysis of Limit Cycles Due to Sign-Magnitude Truncation in Multiplication in Recursive Digital Filters," *Proc. 5th Ansilomar Conf. Circuits Systems*, 1971.
11. R. B. Blackman, *Linear Data-Smoothing and Prediction in Theory and Practice*,

- Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.
12. P. M. Ebert, J. E. Mazo, and M. C. Taylor, "Overflow Oscillations in Digital Filters," *Bell System Tech. J.*, Vol. 48, 1969, pp. 2999-3020.
13. L. B. Jackson, "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," *Bell System Tech. J.*, Vol. 49, 1970, pp. 159-184.
14. L. B. Jackson, "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade of Parallel Form," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, June 1970, pp. 107-122.
15. E. P. F. Kan and J. K. Aggarwal, "Error Analysis of Digital Filter Employing Floating-Point Arithmetic," *IEEE Trans. Circuit Theory*, Vol. CT-18, Nov. 1971, pp. 678-686.
16. B. Liu and T. Kaneko, "Error Analysis of Digital Filters Realized with Floating-Point Arithmetic," *Proc. IEEE*, Vol. 57, Oct. 1969, pp. 1735-1747.
17. I. W. Sandberg, "Floating-Point-Roundoff Accumulation in Digital Filter Realization," *Bell System Tech. J.*, Vol. 46, Oct. 1967, pp. 1775-1791.
18. C. Weinstein and A. V. Oppenheim, "A Comparison of Roundoff Noise in Floating Point and Fixed Point Digital Filter Realizations," *Proc. IEEE (Lett.)*, Vol. 57, June 1969, pp. 1181-1183.
19. C. J. Weinstein, "Quantization Effects in Digital Filters," *MIT Lincoln Lab. Tech. Rept. 468, ASTIA DOC. DDC AD-706862*, Nov. 21, 1969.
20. D. S. K. Chan and L. R. Rabiner, "Analysis of Quantization Errors in the Direct Form for Finite Impulse Response Digital Filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, No. 4, Aug. 1973, pp. 354-366.
21. D. S. K. Chan and L. R. Rabiner, "Theory of Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters," *Bell System Tech. J.*, Vol. 52, No. 3, Mar. 1973, pp. 329-345.
22. D. S. K. Chan and L. R. Rabiner, "An Algorithm for Minimizing Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters," *Bell System Tech. J.*, Vol. 52, No. 3, Mar. 1973, pp. 347-385.
23. P. D. Welch, "A Fixed-Point Fast Fourier Transform Error Analysis," *IEEE Trans. Audio Electroacoust.*, Vol. AU-17, June 1969, pp. 153-157.
24. W. M. Gentleman and G. Sande, "Fast Fourier Transforms—for Fun and Profit," in *Proc. 1966 Fall Joint Computer Conf., AFIPS Conf. Proc.*, Vol. 29, pp. 563-578, Spartan Books, Washington, D.C., 1966.
25. T. Kaneko and B. Liu, "Accumulation of Roundoff Error in Fast Fourier Transforms," *J. Assoc. Comput. Mach.*, Vol. 17, Oct. 1970, pp. 637-654.
26. C. J. Weinstein, "Roundoff Noise in Floating Point Fast Fourier Transform Computation," *IEEE Trans. Audio Electroacoust.*, Vol. AU-17, Sept. 1969, pp. 209-215.
27. B. Liu, "Effect of Finite Word Length on the Accuracy of Digital Filters—A Review," *IEEE Trans. Circuit Theory*, Vol. CT-18, Nov. 1971, pp. 670-677.
28. A. V. Oppenheim and C. J. Weinstein, "Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform," *Proc. IEEE*, Aug. 1972, pp. 957-976.
29. A. V. Oppenheim, "Realization of Digital Filters Using Block-Floating-Point Arithmetic," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, Jan. 1970, pp. 130-136.

PROBLEMI

1. In questo problema desideriamo considerare alcune delle proprietà delle varie rappresentazioni dei numeri discusse nel testo. Si consideri pertanto un numero x tale che $|x| < 1$ e si assuma anche che $|x|$ sia rappresentabile con una frazione binaria con b bit alla destra della virgola binaria. Per convenienza introduciamo

la notazione \simeq col significato di «è rappresentato da». Pertanto, nel caso del modulo e segno avremo

$$x \simeq \begin{cases} |x|, & x \geq 0 \\ 1 + |x|, & x \leq 0 \end{cases}$$

Per il complemento a uno

$$x \simeq \begin{cases} |x|, & x \geq 0 \\ 2 - 2^{-b} - |x|, & x \leq 0 \end{cases}$$

Per il complemento a due

$$x \simeq \begin{cases} |x|, & x \geq 0 \\ 2 - |x|, & x < 0 \end{cases}$$

- (a) In ciascuno di questi casi, x è rappresentato da un numero binario di $(b+1)$ bit. Dimostrare che in tutti e tre i casi il bit immediatamente alla sinistra della virgola binaria (il bit segno) è 0 per $x > 0$ ed 1 per $x < 0$.
 - (b) Dimostrare che il seguente algoritmo è sufficiente per ottenere la rappresentazione in complemento a uno di un numero negativo x a partire dal suo modulo $|x|$: «Ottieni il negativo di $|x|$ cambiando ogni 1 in 0 ed ogni 0 in 1, incluso il bit segno».
 - (c) Dimostrare che i seguenti algoritmi servono per ottenere la rappresentazione in complemento a due di un numero negativo x a partire dal suo modulo $|x|$.
 - (1) «Trova il complemento a uno e somma 2^{-b} ».
 - (2) «Partendo da destra esamina i bit di $|x|$ uno alla volta. Per ogni 0 in $|x|$ metti uno 0 in x . Quando incontri il primo 1 in $|x|$ metti un 1 in x . Di là in poi cambia ogni 0 in 1 ed ogni 1 in 0, per tutti i bit incluso il bit segno».
2. La rappresentazione in complemento a due comporta delle semplificazioni molto interessanti dei processi aritmetici. Per dimostrarlo prenderemo in esame alcuni dettagli dell'addizione in complemento a due. Come nel probl. 9.1, useremo la notazione \simeq col significato di «è rappresentato da». Un numero x ha dunque la rappresentazione in complemento a due

$$x \simeq \begin{cases} |x|, & x \geq 0 \\ 2 - |x|, & x < 0 \end{cases}$$

dove $|x| < 1$ e il numero di bit usati per la rappresentazione di x è $(b+1)$. La somma in complemento a due si effettua nel modo seguente:

- (1) Tutti i numeri sono trattati come numeri binari *senza segno* di $(b+1)$ bit.
- (2) La somma è la semplice somma binaria.
- (3) I riporti oltre il bit segno sono ignorati; cioè, se la somma è maggiore di 2 il riporto viene ignorato. Pertanto la somma è realizzata in modulo 2.
 - (a) Usando le notazioni e definizioni date sopra, scrivere le espressioni complete per la somma in complemento a due di due numeri x_1 e x_2 dove $|x_1|$ e $|x_2| < 1$. Si considerino tutti i casi possibili; si noti, cioè, che x_1 e x_2 possono essere sia $+$ che $-$ ed anche che $|x_1|$ può essere sia maggiore che minore di $|x_2|$.
 - (b) Si osservi che quando si sommano due numeri dello stesso segno, l'ampiezza risultante può essere maggiore di 1. Questa condizione si chiama *saturazione*. Dimostrare che il verificarsi della saturazione risulta indicato dal fatto che la somma di due numeri dello stesso segno ha segno opposto.
 - (c) Dimostrare che la somma in complemento a due di x_1 e x_2 è equivalente a $f[x_1 + x_2]$ dove $f[\]$ è rappresentata in fig. P9.2.
 - (d) Siano $x_1 = 5/8$, $x_2 = 3/4$ e $x_3 = -1/2$. Trovare per ciascun numero le rappresentazioni in complemento a due e sommare tali rappresentazioni nella sequenza $(x_1 + x_2) + x_3$. Notare che nella somma $(x_1 + x_2)$ si ha saturazione, ma il risultato finale è corretto. Si dimostri che in generale si può verificare

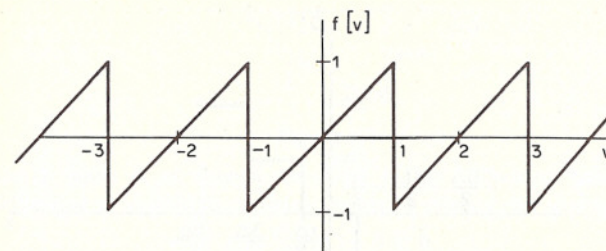


Fig. P9.2

un numero qualsiasi di saturazioni nel processo di accumulazione della somma di tre o più numeri in complemento a due, e tuttavia il risultato sarà corretto se la somma vera ha modulo minore di uno.

3. In una rappresentazione dei numeri in virgola mobile, il numero è rappresentato nella forma

$$F = 2^c M, \quad \frac{1}{2} \leq |M| < 1$$

dove c è comunemente chiamato la caratteristica ed M la mantissa. Poiché $1/2 \leq |M| < 1$, si tratta di una frazione in virgola fissa. I numeri negativi in virgola mobile possono essere rappresentati rappresentando la mantissa come una frazione in virgola fissa nella notazione in modulo e segno, o in complemento a uno, o in complemento a due.

Si consideri un numero in virgola mobile F che deve essere quantizzato quantizzando la mantissa in b bit, segno escluso, così che il valore del bit meno significativo nella mantissa è 2^{-b} .

Si indichi con $Q[F]$ il valore quantizzato. È conveniente rappresentare $Q[F]$ come $Q[F] = F(1 + \epsilon)$, in modo che l'errore $E = Q[F] - F$ sia

$$E = \epsilon F$$

- (a) Nell'ipotesi che F sia un numero positivo, dimostrare che nel caso dell'arrotondamento si ha $-2^{-b} < \epsilon \leq 2^{-b}$ e che nel caso del troncamento si ha $-2 \cdot 2^{-b} < \epsilon \leq 0$.
 - (b) Nell'ipotesi che F sia un numero negativo si determinino una limitazione superiore e una inferiore su ϵ nei casi di: (1) arrotondamento; (2) troncamento in modulo e segno; (3) troncamento in complemento a uno; (4) troncamento in complemento a due.
4. Per poter elaborare una sequenza con un calcolatore numerico, dobbiamo quantizzare l'ampiezza della sequenza in un insieme di livelli discreti. La quantizzazione può rappresentarsi in termini di passaggio della sequenza di ingresso $x(n)$ attraverso un quantizzatore $Q(x)$ che ha una relazione ingresso-uscita come quella illustrata nella fig. P9.4-1.
- Se l'intervallo di quantizzazione Δ è piccolo rispetto ai cambiamenti di livello della sequenza di ingresso, si può supporre che l'uscita del quantizzatore $y(n)$ sia della forma

$$y(n) = x(n) + e(n)$$

dove $e(n) = Q[x(n)] - x(n)$ ed $e(n)$ è un processo casuale stazionario con densità di probabilità del primo ordine uniforme tra $-\Delta/2$ e $\Delta/2$, con campioni scorrelati tra loro e con $x(n)$, così che $E[e(n)x(m)] = 0$ per tutti gli m ed n .

Nell'ipotesi che $x(n)$ sia un rumore bianco con media zero e varianza σ_x^2 .

- (a) Trovare la media, la varianza e la sequenza di autocorrelazione di $e(n)$.
- (b) Quanto vale il rapporto segnale-rumore di quantizzazione σ_x^2/σ_e^2 ?
- (c) Il segnale quantizzato $y(n)$ deve essere filtrato con un filtro numerico con risposta all'impulso $h(n) = 1/2[a^n + (-a)^n]u(n)$. Si determini la varianza

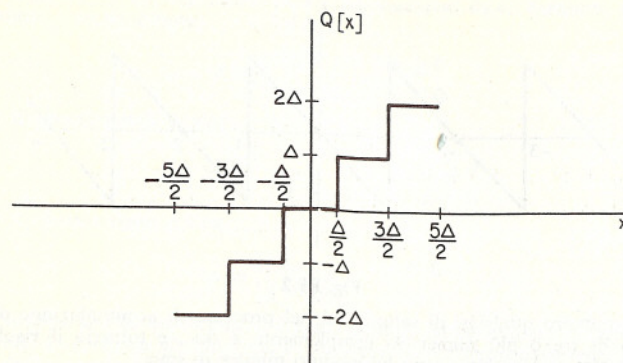


Fig. P9.4-1

del rumore prodotto all'uscita dovuto al rumore di quantizzazione in ingresso e si determini il rapporto segnale-rumore in uscita.

In alcuni casi può convenire usare intervalli di quantizzazione non lineari come, per esempio, intervalli di quantizzazione spaziali logaritmicamente. Ciò si può fare applicando la quantizzazione uniforme al logaritmo dell'ingresso come illustrato nella fig. P9.4-2

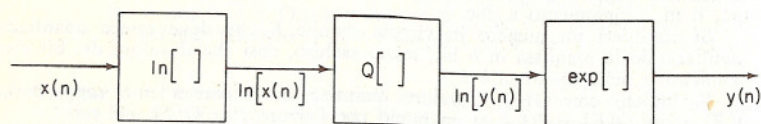


Fig. P9.4-2

dove $Q[]$ è un quantizzatore uniforme come definito sopra. In questo caso, se facciamo l'ipotesi che Δ sia piccolo rispetto alle variazioni nella sequenza $\ln[x(n)]$, si può assumere che l'uscita del quantizzatore sia

$$\ln[y(n)] = \ln[x(n)] + e(n)$$

Pertanto

$$y(n) = x(n) \cdot \exp[e(n)]$$

Per piccoli valori di $e(n)$ si può approssimare $\exp[e(n)]$ con $(1 + e(n))$, così che

$$y(n) \approx x(n)[1 + e(n)] = x(n) + f(n)$$

Questa equazione verrà usata per descrivere l'effetto della quantizzazione logaritmica. Supporremo che $e(n)$ sia un processo casuale stazionario, scorrelato tra campione e campione, indipendente dal segnale $x(n)$ e con densità di probabilità del primo ordine uniforme tra $-\Delta/2$ e $+\Delta/2$.

- (d) Si determinino media, varianza e sequenza di autocorrelazione del rumore additivo $f(n)$ definito sopra.
- (e) Quanto vale il rapporto segnale-rumore di quantizzazione σ_x^2/σ_f^2 ? Si noti che quest'ultimo è in questo caso indipendente da σ_x^2 per cui, entro i limiti delle nostre ipotesi, il rapporto segnale-rumore di quantizzazione è indipendente dal livello del segnale di ingresso, mentre nel caso della quantizzazione lineare esso dipende direttamente da σ_x^2 .
- (f) Il segnale quantizzato $y(n)$ deve essere filtrato per mezzo di un filtro numerico con risposta all'impulso $h(n) = 1/2[a^n + (-a)^n]u(n)$. Si determini la varianza del rumore prodotto all'uscita dovuto al rumore di quantizzazione in ingresso e si determini il rapporto segnale-rumore in uscita.

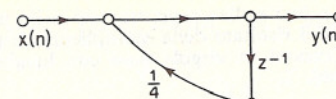


Fig. P9.5-1

5. Il grafo di flusso di un sistema del primo ordine è illustrato nella fig. P9.5-1. (a) Nell'ipotesi di aritmetica perfetta, trovare la risposta del sistema all'ingresso

$$x(n) = \begin{cases} \frac{1}{2}, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

Come è la risposta del sistema per n grande?

Si supponga ora che il sistema debba essere realizzato con l'aritmetica in virgola fissa, e che il coefficiente e tutte le variabili della rete siano rappresentati in modulo e segno con registri di 5 bit. Tutti i numeri devono, cioè, considerarsi frazioni con segno della forma

s	a	b	c	d
---	---	---	---	---

s = bit segno

$$\text{valore del registro} = a \times 2^{-1} + b \times 2^{-2} + c \times 2^{-3} + d \times 2^{-4}$$

dove a, b, c e d sono 0 o 1.

Il risultato di ogni moltiplicazione viene troncato; si conservano, cioè, solo il segno e i quattro bit più significativi.

- (b) Si calcoli la risposta del sistema quantizzato all'ingresso della parte (a) e si disegnino le risposte dei due sistemi, quantizzato e non, per $0 \leq n \leq 5$. Come vanno le due risposte l'una rispetto all'altra per n grande?
- (c) Si consideri ora il sistema illustrato nella fig. P9.5-2, dove

$$x(n) = \begin{cases} \frac{1}{2}(-1)^n, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

Si ripetano le parti (a) e (b) per questo nuovo sistema e nuovo ingresso.

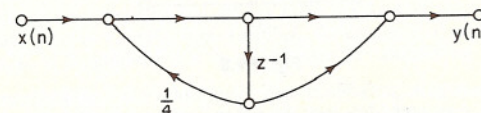


Fig. P9.5-2

6. Si consideri un sistema del primo ordine della forma

$$y(n) = \alpha y(n-1) + x(n)$$

Si supponga che tutte le variabili e i coefficienti siano rappresentati in modulo e segno e che i risultati delle moltiplicazioni siano troncati. Pertanto l'effettiva equazione alle differenze è

$$w(n) = Q[\alpha w(n-1)] + x(n)$$

dove $Q[]$ rappresenta il troncamento in modulo e segno.

Considerare la possibilità di un ciclo limite per ingresso zero della forma $|w(n)| = |w(n-1)|$ per tutti gli n . Dimostrare che se il sistema ideale è stabile non può esistere ciclo limite per ingresso zero. Verificare se vale lo stesso risultato nel caso del troncamento in complemento a due.

7. Si consideri il filtro del primo ordine rappresentato nella fig. P9.7. Il quantizzatore $Q[\]$ sta a indicare che il risultato della moltiplicazione per α viene arrotondato. Tutti i numeri sono frazioni in virgola fissa con lunghezza di parola di b bit (escludendo il bit segno).

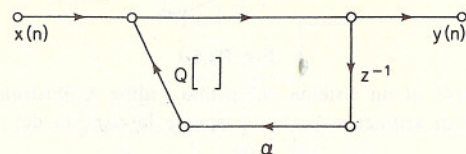


Fig. P9.7

L'ingresso è zero, ma al filtro è associata la condizione iniziale $y(-1) = A$. Per via del quantizzatore esiste un insieme di valori di A , chiamato la banda morta, per il quale il valore effettivo del coefficiente α è $+1$ o -1 , ovvero $|Q(\alpha A)| = A$. Una volta che l'uscita cada in quell'insieme, essa oscillerà o rimarrà costante a seconda che il valore effettivo del coefficiente è positivo o negativo.

- Determinare in termini di α e b l'insieme dei valori di A corrispondenti alla banda morta.
 - Nel caso in cui $b = 6$ bit e $A = 1/16$, disegnare $y(n)$ per $\alpha = +15/16$ e $\alpha = -15/16$.
 - Nel caso in cui $b = 6$ ed $A = 1/2$, disegnare $y(n)$ per $\alpha = -15/16$.
8. Si consideri il sistema del secondo ordine illustrato nella fig. P9.8. Il filtro viene realizzato con l'aritmetica in virgola fissa e i risultati di tutte le moltiplicazioni vengono arrotondati. Tutti i numeri sono frazioni in virgola fissa con lunghezza di parola di b bit. Analogamente al caso del probl. 7 di questo capitolo, esiste una zona di banda morta per $y(n)$ per cui il valore « effettivo » del coefficiente $-r^2$ è -1 . Quando $y(n)$ è interno a questo insieme di valori, i poli effettivi si trovano sul circolo unitario.

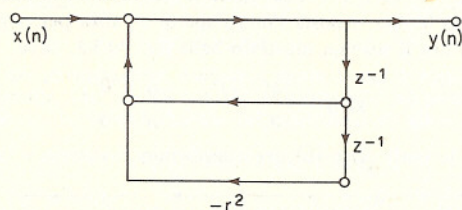


Fig. P9.8

Siano $x(n) = 0$, $y(-1) = A$, $y(-2) = 0$, $A \neq 0$.

- Trovare la banda morta per A , cioè, i valori di A per cui $Q[-r^2 A] = -A$.
 - Ottenendo una limitazione inferiore su A , trovare l'insieme di valori di r per cui è possibile avere una banda morta.
9. Si consideri il sistema del secondo ordine illustrato nella fig. P9.9-1.

La funzione $f[\]$ è illustrata nella fig. P9.9-2 e, come si è già visto nel precedente probl. 2, rappresenta la somma in complemento a due. In altre parole, la funzione $f[\]$ tiene conto di possibili saturazioni nel formare la somma

$$ay(n-1) + by(n-2) + x(n)$$

Un semplice esempio mostra la possibilità di esistenza di cicli limite per ingresso zero. Trascureremo gli arrotondamenti nei prodotti $ay(n-1)$ e $by(n-2)$.

- Si determini innanzitutto l'insieme dei valori di a e b che assicurano la stabilità del sistema quando non si ha saturazione. Rappresentare la regione del piano a - b corrispondente alla stabilità in condizioni di linearità.

$$y(n) = f[x(n) + ay(n-1) + by(n-2)]$$

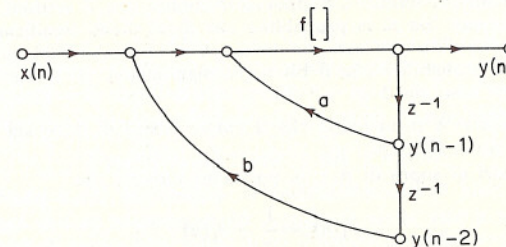


Fig. P9.9-1

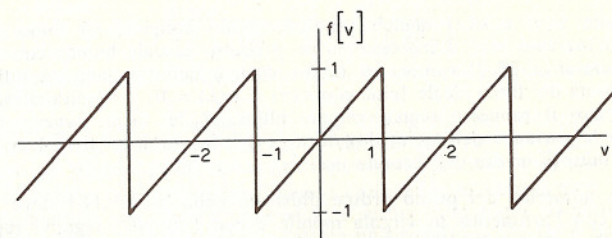


Fig. P9.9-2

- Si supponga che sia $x(n) = 0$. Quali condizioni su a e b assicurano che non si abbia saturazione, e, cioè, $y(n) < 1$? Tratteggiare la corrispondente regione del piano a - b .
 - Si consideri la possibilità che, con $x(n) = 0$, sia $y(n) = y_0 > 0$ per tutti gli n . Quale è il valore di y_0 quando a e b soddisfano le condizioni di stabilità lineare?
 - Quali valori di a e b sono compatibili con la condizione di stabilità e con la condizione $0 < y_0 < 1$?
 - Si consideri ora la possibilità di un ciclo limite per ingresso zero di periodo due, cioè, $y(n) = (-1)^n y_0$ per tutti gli n . Quale è il valore possibile di y_0 , con $0 < y_0 < 1$, quando a e b soddisfano la condizione di stabilità lineare?
10. Quando si realizzano i filtri numerici con lunghezza di parola finita per i coefficienti, non è possibile collocare i poli e gli zeri del filtro con precisione arbitraria. Un modo per compensare questa limitazione è quello di introdurre del « jitter » casuale per ogni coefficiente in modo tale che il valore medio del coefficiente è quello desiderato.

Si consideri la realizzazione della rete numerica passa-tutto illustrata nella fig. P9.10, dove $k_2 = -1/k_1$. Nel corso di questo problema faremo l'ipotesi che non ci sia rumore di arrotondamento di tipo aritmetico ma che il solo rumore introdotto sia dovuto al « jitter » casuale aggiunto ai coefficienti.

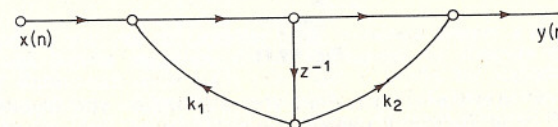


Fig. P9.10

Si vuole che il valore medio del coefficiente k_1 sia $1/\pi$ e il valore medio del coefficiente k_2 sia $-\pi$. I coefficienti sono tutti rappresentati in modulo e segno e la lunghezza di parola è scelta in modo che alla parte frazionaria del coefficiente siano riservati sette bit. Supponiamo inoltre che il settimo bit meno significativo sia casuale. Sia p_1 la probabilità che il bit meno significativo in k_1 sia 1, col che la probabilità che esso sia invece 0 è $(1-p_1)$. Analogamente poniamo uguale a p_2 la probabilità che il bit meno significativo di k_2 sia 1, e $(1-p_2)$ la probabilità che esso sia 0.

(a) Determinare p_1 e p_2 in modo che il valore atteso di $k_1(n)$ sia $1/\pi$ e il valore atteso di $k_2(n)$ sia $-\pi$.

(b) Con i valori di sopra di p_1 e p_2 possiamo scrivere che

$$k_1(n) = \frac{1}{\pi} + \varepsilon_1(n)$$

$$k_2(n) = -\pi + \varepsilon_2(n)$$

Siano $\varepsilon_1(n)$ e $\varepsilon_2(n)$ bianchi, statisticamente indipendenti l'uno dall'altro e dall'ingresso $x(n)$. L'ingresso sia un processo casuale bianco con media zero e varianza σ_x^2 . Il rumore di uscita, $f(n)$, è definito come la differenza tra l'uscita del filtro ideale [cioè con $\varepsilon_1(n) = \varepsilon_2(n) = 0$] e l'uscita effettiva. Determinare il rapporto segnale-rumore all'uscita del filtro, ovvero, il rapporto tra la varianza dell'uscita dovuta a $x(n)$ e la varianza di $f(n)$. Si ignorino i termini di errore del secondo ordine.

11. Il filtro numerico del primo ordine illustrato nella fig. P9.11-1 deve essere realizzato con l'aritmetica in virgola mobile e con i numeri negativi rappresentati il modulo e segno. Per il coefficiente α il numero di bit assegnati alla mantissa, segno escluso, è b .

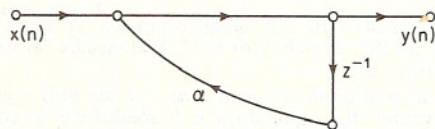


Fig. P9.11-1

Quanto alla caratteristica, faremo l'ipotesi che il numero di bit disponibili sia illimitato. La mantissa risultante dalla moltiplicazione è troncata a b bit. Per semplificare l'analisi faremo l'ipotesi che la mantissa risultante dalla addizione non sia troncata.

L'errore dovuto al troncamento del prodotto sarà rappresentato da una sorgente di rumore additivo $e(n)$, come mostrato nella fig. P9.11-2

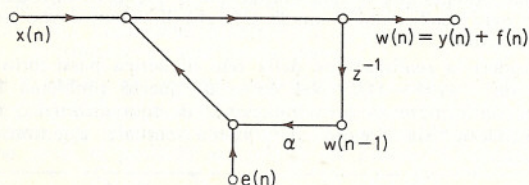


Fig. P9.11-2

con $e(n) = \alpha \cdot \varepsilon(n) y(n-1)$. L'uscita è $y(n) + f(n)$, dove $y(n)$ rappresenta l'uscita che si avrebbe in assenza di rumore di quantizzazione, cioè con $e(n)$ zero.

Circa $\varepsilon(n)$ si fanno le seguenti ipotesi:

- (1) $\varepsilon(n)$ è un processo casuale stazionario con densità di probabilità del primo ordine, $p(\varepsilon)$, uniforme, data da

$$p(\varepsilon) = \begin{cases} \frac{1}{2} \cdot 2^b, & 0 < \varepsilon < 2 \cdot 2^{-b} \\ 0, & \text{altrove} \end{cases}$$

- (2) $\varepsilon(n)$ è statisticamente indipendente da $x(n)$ e $y(n)$.

- (3) $E[\varepsilon(n) \varepsilon(m)] = E[\varepsilon(n)] \cdot E[\varepsilon(m)]$ per $n \neq m$.

L'ingresso $x(n)$ è un processo casuale con funzione di autocorrelazione

$$\phi_{xx}(n) = E[x(r)x(r+n)] = \sigma_x^2 \delta(n)$$

- (a) Determinare $\phi_{yy}(n)$, la funzione di autocorrelazione di $y(n)$.

- (b) Indicando con σ_y^2 la varianza del segnale di uscita $y(n)$ e con σ_f^2 la varianza del rumore di uscita $f(n)$, determinare il rapporto rumore-segnale σ_f^2/σ_y^2 .

12. Si vuole realizzare un filtro numerico avente come funzione di trasferimento

$$H(z) = \frac{1 - \frac{1}{2}z^{-1}}{(1 - \frac{1}{2}z^{-1})(1 + \frac{1}{2}z^{-1})}$$

- (a) Disegnare un grafo di flusso della rete numerica che realizza il filtro in forma canonica.

- (b) Disegnare i grafi di flusso delle reti numeriche che realizzano il filtro in ogni possibile struttura a cascata contenente solo reti del primo ordine (esistono sei possibilità, di cui quattro contenenti il numero minimo di ritardi).

- (c) Il filtro deve essere realizzato usando l'aritmetica in virgola fissa. I dati sono rappresentati con frazioni in modulo e segno e lunghezza di parola di b bit, segno escluso. Il moltiplicatore calcola un prodotto di $2b$ bit e poi arrotonda il risultato ai b bit più significativi. Trascurando il problema delle saturazioni, si determini, per ognuna delle sette reti di (a) e (b), la varianza del rumore in uscita dovuto all'arrotondamento del moltiplicatore. Osservando i grafi di flusso, si dovrebbe essere capaci di vedere che alcuni di essi danno luogo alla stessa risposta. Indicare quale delle configurazioni dà luogo al più basso rumore di uscita.

- (d) Se da un lato l'aritmetica in virgola fissa non introduce rumore di arrotondamento nelle addizioni, dall'altro essa comporta una limitata dinamica, per cui la scala dell'ingresso deve essere corretta in modo tale che nel filtro nessun valore di segnale ecceda la lunghezza del registro. Partendo dalla somma di convoluzione si può trovare una limitazione per l'uscita di un sistema lineare in termini del massimo valore dell'ingresso e della somma dei valori assoluti della risposta all'impulso, come indicato nelle seguenti espressioni

$$y(n) = \sum_{k=-\infty}^{\infty} x(n-k)h(k)$$

$$|y(n)| \leq \sum_{k=-\infty}^{\infty} |x(n-k)| |h(k)|$$

Se x_{\max} è il massimo valore dell'ingresso e y_{\max} il massimo valore dell'uscita, si ha

$$y_{\max} \leq x_{\max} \sum_{k=-\infty}^{\infty} |h(k)|$$

Va notato che, per un qualsiasi filtro, possiamo sempre trovare un ingresso tale che l'uscita raggiunge il valore massimo. In ciascuna delle configurazioni determinate nelle parti (a) e (b), il livello del segnale raggiunge un valore massimo in qualche nodo della rete, che non coincide necessariamente con l'uscita. Usando le espressioni di sopra, determinare in funzione di x_{\max} il valore massimo che può esistere in qualsiasi punto della rete. Effettuare questa analisi per ciascuna delle sette reti delle parti (a) e (b).

- (e) Facciamo l'ipotesi che l'ingresso sia una sequenza rumore bianco, uniformemente distribuito in ampiezza tra $-x_{\max}$ e $+x_{\max}$. Per ciascuno dei filtri delle parti (a) e (b) e sulla base delle risposte alla parte (d), potremo scegliere un valore per x_{\max} tale che il massimo livello di segnale nel filtro sia 1 (poiché stiamo lavorando con frazioni in virgola fissa, questo è effettivamente il valore più grande che può stare in un registro). Possiamo quindi definire un rapporto rumore-segnale all'uscita di ogni configurazione di filtro come il rapporto tra la varianza del rumore in uscita (determinata come in (c)) e il valore quadratico medio del segnale di uscita. Si determini il rapporto rumore-segnale per ognuna delle sette configurazioni. Indicare quale delle configurazioni dà luogo al rapporto rumore-segnale più basso.

13. Il sistema definito dalla equazione alle differenze

$$y(n) - \alpha y(n-1) = x(n) - \frac{1}{\alpha} x(n-1)$$

è un sistema passa-tutto, tale, cioè, che il modulo della sua risposta in frequenza è costante, indipendente dalla frequenza. Vogliamo, per tale sistema, confrontare l'effetto dell'arrotondamento aritmetico per una realizzazione in virgola fissa e una in virgola mobile. Si considerino tutti i numeri in virgola fissa come frazioni, per cui i valori sono tra -1 e $+1$, e con lunghezza di registro assegnata di b bit, segno escluso. Per il caso in virgola mobile, sia t il numero di bit della mantissa, segno escluso.

Si faccia l'ipotesi che α sia reale e tale che $1/2 < \alpha < 1$, e che l'ingresso al filtro $x(n)$ sia un processo casuale bianco, con ampiezza uniformemente distribuita fra $-x_0$ e $+x_0$.

- (a) Il massimo livello del segnale di ingresso deve essere abbastanza piccolo da non provocare saturazione nella realizzazione del filtro in virgola fissa. Tenendo conto di ciò, determinare per la realizzazione in virgola fissa, sia per la forma diretta che per quella canonica, il rapporto rumore-segnale in uscita, ovvero il rapporto tra la varianza dell'uscita dovuta al rumore di arrotondamento e la varianza dell'uscita dovuta a $x(n)$.
- (b) Determinare il rapporto rumore-segnale di uscita per la forma canonica del filtro realizzato con l'aritmetica in virgola mobile.
- (c) Si supponga α prossimo a uno, per cui $\alpha = 1 - \delta$, $\delta \ll 1$. Esprimere in termini di δ i risultati ottenuti in (a) e (b), facendo ragionevoli approssimazioni.
- (d) La realizzazione in virgola mobile di un filtro richiede ulteriori bit per la caratteristica. Sia B il numero di bit per la caratteristica nella parola in virgola mobile. Se, per ciascuna delle due realizzazioni in virgola fissa (forma canonica e forma diretta), è $b = B + t$, si determini B come funzione di δ in modo che i rapporti rumore-segnale per la virgola fissa e la virgola mobile siano uguali.
14. Abbiamo finora discusso le prestazioni dei filtri numerici (in termini di rumore) nelle realizzazioni con l'aritmetica in virgola fissa e in virgola mobile. In questo problema prenderemo in esame una alternativa, chiamata *aritmetica in virgola mobile a blocchi* (v. [29]). Per illustrare il procedimento, si consideri un filtro del primo ordine definito dall'equazione alle differenze

$$y(n) = x(n) + \alpha y(n-1)$$

corrispondente alla rete mostrata nella fig. P9.14-1.

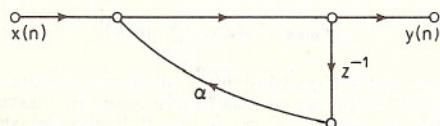


Fig. P9.14-1

Se nel filtro la moltiplicazione e la somma sono realizzate in virgola fissa, non si utilizza pienamente in ogni iterazione la lunghezza del registro, in quanto

l'ingresso deve essere ridotto di scala affinché il massimo valore di uscita sia 1. La virgola mobile a blocchi corrisponde a moltiplicare $x(n)$ e $y(n-1)$ per un guadagno $A(n)$ in modo da normalizzarli congiuntamente, e dividere quindi $y(n)$ per questo stesso guadagno, ovvero

$$y(n) = \frac{1}{A(n)} [x(n)A(n) + \alpha y(n-1)A(n)]$$

dove $A(n)$ ovviamente cambia da iterazione a iterazione.

Nel corso di questo problema supporremo che $x(n)$ sia un processo casuale bianco con media zero e densità di probabilità uniforme tra $-x_0$ e $+x_0$.

- (a) Dimostrare che la rete di fig. P9.14-2 è equivalente alla rete di fig. P9.14-1 se si trascurano gli effetti di quantizzazione.

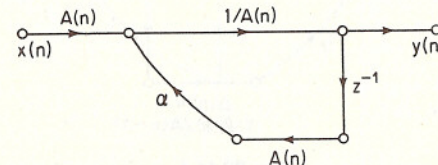


Fig. P9.14-2

- (b) La moltiplicazione e la somma sono effettuate con l'aritmetica in virgola fissa. Il guadagno $A(n)$ è scelto come

$$A(n) = 2^{c(n)}, \quad c(n) \geq 0$$

per cui esso corrisponde a una traslazione a sinistra dei registri. Poiché $A(n)$ è scelto in modo da normalizzare congiuntamente $x(n)$ e $y(n-1)$, avremo

$$\frac{1}{2} \leq A(n) \max\{|x(n)|, |y(n-1)|\} < 1$$

dove $\max\{|x(n)|, |y(n-1)|\}$ è il valore più grande tra $|x(n)|$ e $|y(n-1)|$ per ogni n .

Dobbiamo, perciò, imporre che sia

$$\frac{1}{2} \leq A(n) < \frac{1}{\max\{|x(n)|, |y(n-1)|\}} \quad (\text{P9.14-1})$$

Poiché la moltiplicazione di un registro per una potenza positiva di due non comporta rumore di arrotondamento, le sorgenti di rumore introdotte, dovute all'arrotondamento del moltiplicatore, sono mostrate nella fig. P9.14-3.

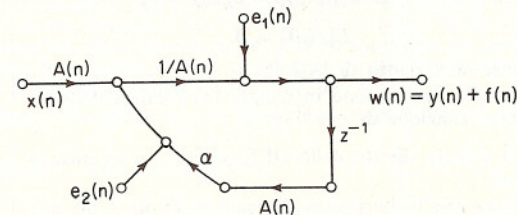


Fig. P9.14-3

Assumendo che $e_1(n)$ ed $e_2(n)$ siano uniformemente distribuite in $\pm 0.5 \cdot 2^{-b}$ nonché scorrelate tra loro e con $x(n)$ [e di conseguenza con $A(n)$], determinare la varianza del rumore di uscita $f(n)$ per la rete di sopra. Esprimere la risposta in termini di $k^2 = E[(1/A(n))^2]$. Dimostrare che la varianza del rumore in uscita dovuto all'arrotondamento aritmetico per la rete di figura

P9.14-3 è sempre maggiore di quanto si otterrebbe se il filtro del primo ordine fosse realizzato direttamente, come mostrato nella fig. P9.14-2.

- (c) La rete di fig. P9.14-3 può essere modificata in una forma equivalente, mostrata nella fig. P9.14-4. Siano $e_1(n)$, $e_2(n)$ ed $e_3(n)$ i rumori introdotti dall'arrotondamento nelle moltiplicazioni per $[1/A(n)]$, $\Delta(n)$ e α , rispettivamente. Si supponga, anche qui, che tali rumori siano uniformemente distribuiti in $\pm 0.5 \cdot 2^{-b}$ e indipendenti l'uno dall'altro nonché da $x(n)$ e $y(n)$. Si determini la varianza del rumore di uscita complessivo dovuto a queste tre sorgenti di rumore. Esprimere la risposta in termini di $k^2 = E[(1/A(n))^2]$.

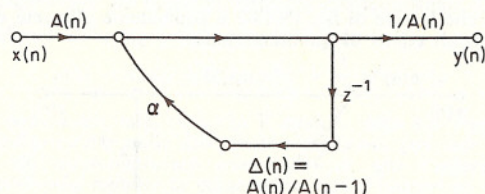


Fig. P9.14-4

- (d) Si assuma $\alpha = 1 - \delta$, $\delta \ll 1$ (caso di guadagno elevato). Per questo valore di α , si ammette che il valore assoluto dell'uscita, $|y(n-1)|$, sia maggiore del valore assoluto dell'ingresso $|x(n)|$. Su questa base e usando la (P9.14-1), dimostrare che una limitazione superiore per la quantità $k^2 = E[(1/A(n))^2]$ è $k^2 \leq 4\sigma_\delta^2$.
- (e) Per determinare il rapporto rumore-segnale, occorre assicurarsi che il valore massimo di $x(n)$ sia abbastanza piccolo da garantire che sia $|y(n)| < 1$. Si supponga che $x(n)$ sia un processo bianco uniformemente distribuito e che $\alpha = 1 - \delta$, $\delta \ll 1$. Sulla base di queste ipotesi e dei risultati della parte (d), determinare il rapporto rumore-segnale per la rete di fig. P9.14-4. Confrontare questo rapporto rumore-segnale con il corrispondente risultato per l'aritmetica in virgola fissa.

15. Si consideri la DFT

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad 0 \leq k \leq N-1$$

dove $W_N = e^{-j(2\pi/N)}$. Si supponga che i valori della sequenza $x(n)$ siano N valori consecutivi di una sequenza di rumore bianco e cioè

$$E[x(n)x(r)] = \sigma_\delta^2 \delta(n-r)$$

$$E[x(n)] = 0$$

- (a) Determinare la varianza di $|X(k)|^2$.
- (b) Determinare la correlazione incrociata fra i valori della DFT e cioè $E[X(k)X^*(r)]$ come funzione di k e di r .
16. Si consideri il calcolo diretto della DFT usando l'aritmetica in virgola fissa con arrotondamento.
- Si supponga che la lunghezza di registro sia di b bit più il segno e che il rumore di arrotondamento introdotto da ogni moltiplicazione sia indipendente da quello introdotto dalle altre. Nell'ipotesi che $x(n)$ sia reale si determini la varianza del rumore di arrotondamento sia per la parte reale che immaginaria di ogni valore $X(k)$.

17. Nel metodo di calcolo di Goertzel della trasformata di Fourier discreta, si ha

$$X(k) = y_k(r)|_{r=N}$$

dove $y_k(r)$ è l'uscita della rete illustrata nella fig. P9.17. Si consideri la realizzazione dell'algoritmo di Goertzel basato sull'uso dell'aritmetica in virgola fissa con arrotondamento. Si supponga che la lunghezza di registro sia di b bit più il segno e che il rumore di arrotondamento introdotto da una moltiplicazione sia indipendente da quello introdotto da tutte le altre moltiplicazioni.

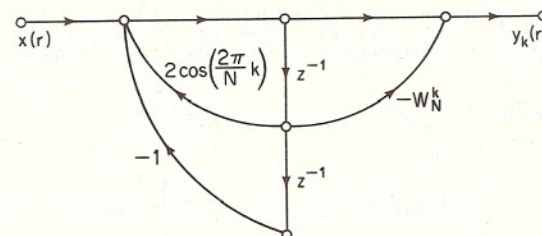


Fig. P9.17

- (a) Nell'ipotesi che $x(r)$ sia reale, dare una rappresentazione mediante grafo di flusso del calcolo con precisione finita delle parti reale e immaginaria di $X(k)$. Si faccia l'ipotesi che la moltiplicazione per ± 1 non introduca rumore di arrotondamento.
- (b) Calcolare la varianza del rumore di arrotondamento sia per la parte reale che per quella immaginaria di ogni valore $X(k)$.
18. Nel par. 9.4.2 abbiamo dimostrato che nella realizzazione in forma diretta e in virgola mobile di un filtro FIR, l'ordine in cui vengono effettuate le moltiplicazioni e le addizioni è un fattore significativo nella determinazione del rapporto rumore-segnale in uscita. In questo problema prenderemo in esame un ordinamento delle addizioni che dà luogo a una significativa riduzione del rapporto rumore-segnale.

Si supponga che $N = 2^v$, dove v è un intero. Decidiamo di calcolare la somma di convoluzione

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k)$$

sommando a due a due i prodotti indicati nella sommatoria, poi sommando ancora a coppie tali somme e così via. Una tale realizzazione è illustrata nella fig. P9.18 per $N = 2^3 = 8$. In questo grafo di flusso, i guadagni tempo-varianti $1 + \epsilon_k(n)$ rappresentano gli errori risultanti dalla quantizzazione delle moltiplicazioni $h(k)x(n-k)$, mentre i guadagni tempo-varianti $1 + \eta_{ij}(n)$ rappresentano gli errori introdotti dalla quantizzazione delle addizioni in virgola mobile. Si fa l'ipotesi che le quantità $\epsilon_k(n)$ e $\eta_{ij}(n)$ siano sequenze di rumore bianco uniformemente distribuite e indipendenti l'una dall'altra.

- (a) Generalizzare la notazione del diagramma precedente al caso generale di $N = 2^v$.
- (b) Esprimere l'uscita $w(n)$ nella forma

$$w(n) = \sum_{k=0}^{N-1} A(n,k)h(k)x(n-k)$$

ovvero esprimere $A(n,k)$ per mezzo delle quantità $1 + \epsilon_k(n)$ e $1 + \eta_{ij}(n)$.

- (c) Dimostrare che

$$E[A(n,k) - 1] = 0$$

$$E[A^2(n,k)] = (1 + \frac{2^{-2b}}{3})^{v+1}$$

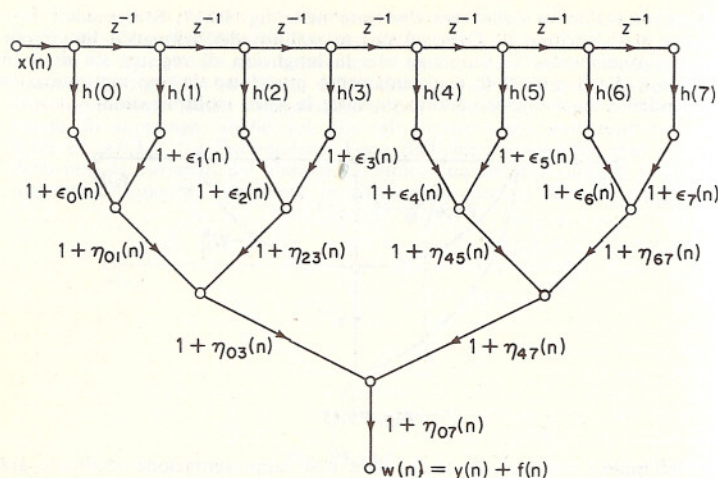


Fig. P9.18

- (d) Ottenere un'espressione per la varianza del rumore di uscita, σ_f^2 , nell'ipotesi che l'ingresso sia un segnale bianco con varianza σ_x^2 e che $(1/3) \cdot 2^{-2b} \ll 1$.
- (e) Dimostrare che il rapporto rumore-segnale in uscita è

$$\frac{\sigma_f^2}{\sigma_y^2} = (\nu + 1) \frac{2^{-2b}}{3}$$

- (f) Confrontare il risultato di (e) con la (9.77).

19. Nel derivare le formule del rapporto rumore-segnale per i calcoli della FFT in virgola fissa, si è supposto ogni nodo di uscita connesso a $(N-1)$ farfalle, ciascuna delle quali contribuisce per $\sigma_F^2 = (1/3) \cdot 2^{-2b}$ alla varianza del rumore in uscita. Tuttavia, quando $W_N^r = \pm 1$ o $\pm j$, le moltiplicazioni possono farsi, in effetti, senza errore. Pertanto i risultati derivati nel par. 9.5.2 possono essere modificati in senso meno pessimistico.
- (a) Nel caso dell'algoritmo di decimazione nel tempo discusso nel paragrafo 9.5.2, si determini per ogni stadio il numero di farfalle che comportano la moltiplicazione per ± 1 o $\pm j$.
- (b) Si usi il risultato di (a) per modificare la stima della varianza del rumore di uscita, (9.87), per il caso in cui tutta la correzione di scala è fatta all'ingresso. Si ricavi inoltre un'espressione modificata, corrispondente alla (9.93), per il rapporto rumore-segnale in uscita.
- (c) Si ripetano (a) e (b) per il caso in cui l'uscita di ogni stadio è attenuata di un fattore $1/2$. In altre parole, si derivino, nell'ipotesi che le moltiplicazioni per ± 1 o $\pm j$ non introducano errore, le espressioni modificate corrispondenti alla (9.95) per la varianza del rumore di uscita, e alla (9.96) per il rapporto rumore-segnale in uscita.
20. Nel derivare le formule del rapporto rumore-segnale nel caso dei calcoli della FFT in virgola mobile, si è supposto che tutti i calcoli di farfalla introducano la stessa quantità di rumore. Tuttavia ciò non è esatto, come si può vedere considerando le moltiplicazioni per ± 1 o $\pm j$.
- (a) Disegnare i grafi di flusso completi, come in fig. 9.22, delle farfalle nel caso della decimazione nel tempo in virgola mobile per $W_N^r = 1$ e per $W_N^r = j$, mostrando tutte le sorgenti di rumore che restano.
- (b) Supponendo che il segnale di ingresso sia bianco, calcolare la varianza del rumore di uscita per una farfalla che contiene $W_N^r = \pm 1$ o $\pm j$.

- (c) Per l'algoritmo a decimazione nel tempo discusso nel paragrafo 9.5.3, determinare per ogni stadio il numero di farfalle che comportano moltiplicazioni per ± 1 o $\pm j$.
- (d) Usando i risultati di (b) e (c), determinare le varianze di rumore medie per le farfalle necessarie al calcolo dell'insieme $(m+1)$ -mo di valori a partire dall'insieme m -mo.
- (e) Usando il risultato di (d), modificare il metodo seguito nel par. 9.5.3 per ottenere un'espressione corrispondente alla (9.99) per la varianza media del rumore in uscita. Usare il risultato per dimostrare che il valore medio del rapporto rumore-segnale in uscita è

$$\frac{E[|F(k)|^2]}{E[|X(k)|^2]} = 2\sigma_x^2 \left[\nu - \frac{3}{2} + \left(\frac{1}{2}\right)^{\nu-1} \right]$$

21. Nel par. 9.5.2 abbiamo presentato un'analisi di rumore per l'algoritmo FFT a decimazione nel tempo della fig. 6.10. Sviluppare un'analisi analoga per l'algoritmo a decimazione in frequenza della fig. 6.18, ottenendo le espressioni per la varianza del rumore in uscita e per il rapporto rumore-segnale, sia nel caso della correzione di scala concentrata in ingresso che in quello della correzione con fattori di scala $1/2$ ad ogni stadio del calcolo.
22. Si consideri il calcolo della DFT mediante coefficienti quantizzati. Sia $X(k)$ la DFT desiderata e $X'(k)$ il risultato che si ottiene invece usando valori quantizzati di W_N^r in un algoritmo FFT di decimazione nel tempo.
- (a) Dimostrare che $X'(k)$ si può esprimere come

$$X'(k) = \sum_{n=0}^{N-1} x(n) \Omega_{nk} = X(k) + F(k)$$

dove

$$\Omega_{nk} = \prod_{i=1}^v (W_N^{r_i} + \delta_i)$$

e

$$\prod_{i=1}^v W_N^{r_i} = W_N^{nk}$$

- (b) Si supponga che i coefficienti $W_N^{r_i}$ siano arrotondati a b bit, segno escluso, e che le parti reale ed immaginaria degli errori nei coefficienti siano scorrelate e uniformemente distribuite. Dimostrare che la variazione delle quantità δ_i è

$$\sigma_{\delta}^2 = \frac{2}{6} \cdot 2^{-2b}$$

- (c) L'errore $F(k)$ si può esprimere come

$$F(k) = X'(k) - X(k) = \sum_{n=0}^{N-1} x(n) (\Omega_{nk} - W_N^{nk})$$

Dimostrare che il fattore $(\Omega_{nk} - W_N^{nk})$ si può esprimere come

$$(\Omega_{nk} - W_N^{nk}) = \sum_{i=1}^v \delta_i \prod_{j=1, j \neq i}^v W_N^{r_j} + \text{termini di ordine superiore}$$

- (d) Trascurando i termini di ordine superiore e supponendo che i δ_i siano mutuamente scorrelati, dimostrare che la varianza di $F(k)$ è

$$\sigma_F^2 = \left(\frac{2\nu}{6}\right) 2^{-2b} \sum_{n=0}^{N-1} |x(n)|^2$$

- (e) Si usi il teorema di Parseval per dimostrare che

$$\frac{\sigma_F^2}{\frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2} = \left(\frac{2\nu}{6}\right) 2^{-2b}$$

10. ELABORAZIONE OMOMORFA DEI SEGNALI

10.0 INTRODUZIONE

Nei capitoli precedenti ci siamo occupati della rappresentazione matematica di segnali e sistemi a tempo discreto, concentrando l'attenzione principalmente sui sistemi lineari invarianti alla traslazione, mentre poco spazio è stato dedicato alle applicazioni. Lo scopo principale di questo capitolo è quello di presentare una classe di tecniche non lineari per l'elaborazione dei segnali. L'analisi di questa classe di tecniche si baserà su gran parte del materiale trattato nei capitoli precedenti. Inoltre le tecniche in questione hanno trovato applicazione in numerosi campi, tra cui il miglioramento di qualità delle immagini, l'analisi della voce e le esplorazioni sismiche. Di conseguenza, l'esame di questa classe di tecniche fornisce l'occasione per illustrare un certo numero di applicazioni di alcuni risultati teorici presentati in questo libro.

La classe di sistemi che discuteremo si basa su una generalizzazione della classe dei sistemi lineari. Abbiamo visto che l'importanza dei sistemi lineari invarianti alla traslazione è dovuta da un lato alla relativa facilità di analisi e caratterizzazione, che porta a rappresentazioni matematiche piuttosto eleganti e potenti, e dall'altro alla possibilità di progettare sistemi lineari invarianti alla traslazione che eseguano una varietà di funzioni utili di elaborazione dei segnali. Ad esempio, se è dato un segnale che è la somma di due segnali le cui trasformate di Fourier occupano bande di frequenza diverse, allora è possibile separare le due componenti con un filtro lineare. Il fatto che i sistemi lineari siano relativamente semplici da analizzare e utili per separare segnali combinati mediante l'operazione di somma è una conseguenza diretta della proprietà di sovrapposizione, che definisce la classe dei sistemi lineari. In base a questa osservazione si è portati a prendere in considerazione classi di sistemi non lineari che obbediscano a un principio di sovrapposizione generalizzato. Tali sistemi sono rappresentati da trasformazioni lineari in senso algebrico tra spazi vettoriali d'ingresso e di uscita e perciò sono stati chiamati *sistemi omomorfi*.

In questo capitolo daremo una breve introduzione alla teoria generale dei sistemi omomorfi e quindi discuteremo in dettaglio due classi di sistemi omomorfi che sono particolarmente adatte per il trattamento di segnali combinati mediante le operazioni di moltiplicazione e convoluzione, essendo questi i due casi in cui la teoria dei sistemi omomorfi è stata applicata

con successo. In questi due esempi vedremo che il principio di sovrapposizione generalizzato può essere usato in maniera del tutto analoga al caso della caratterizzazione dei sistemi lineari. In effetti, si vedrà che il problema di progettare sistemi omomorfi per la moltiplicazione e per la convoluzione si riduce al problema di progettare un sistema lineare.

10.1 SOVRAPPOSIZIONE GENERALIZZATA

Il principio di sovrapposizione enunciato per i sistemi lineari richiede che, se T è la trasformazione eseguita dal sistema, allora per qualsiasi coppia di ingressi $x_1(n)$ e $x_2(n)$ e per qualsiasi scalare c risulti

$$T[x_1(n) + x_2(n)] = T[x_1(n)] + T[x_2(n)] \quad (10.1a)$$

e

$$T[cx_1(n)] = cT[x_1(n)] \quad (10.1b)$$

Per generalizzare questo principio, indichiamo con \square una regola per combinare gli ingressi tra loro (ad es. addizione, moltiplicazione, convoluzione etc.) e con $:$ una regola per combinare gli ingressi con gli scalari. Analogamente, \bigcirc indicherà una regola per combinare le uscite del sistema e $\mathbf{\text{L}}$ una regola per combinarle con gli scalari. Allora, se H rappresenta la trasformazione del sistema, si generalizzano le condizioni (10.1) richiedendo che sia

$$H[x_1(n) \square x_2(n)] = H[x_1(n)] \bigcirc H[x_2(n)] \quad (10.2a)$$

e

$$H[c : x_1(n)] = c \mathbf{\text{L}} H[x_1(n)] \quad (10.2b)$$

Si dice che questi sistemi obbediscono ad un principio di sovrapposizione generalizzato con operazione di ingresso \square e operazione d'uscita \bigcirc . Tali sistemi si rappresentano come in fig. 10.1. È chiaro che i sistemi lineari sono un caso particolare, in cui \square e \bigcirc sono l'addizione e $\mathbf{\text{L}}$ è la moltiplicazione.

La teoria degli spazi vettoriali lineari fornisce il formalismo matematico per la rappresentazione dei sistemi di questa classe. Se interpretiamo gli ingressi e le uscite del sistema come vettori in spazi vettoriali, con le regole \square e \bigcirc che corrispondono alla somma di vettori e le regole $:$ e $\mathbf{\text{L}}$ che corrispondono alla moltiplicazione scalare, allora la trasformazione H del sistema è una trasformazione lineare in senso algebrico dallo spazio vettoriale degli ingressi allo spazio vettoriale delle uscite.

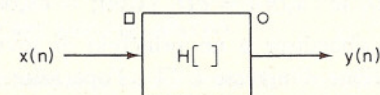


Fig. 10.1 Rappresentazione di un sistema omomorfo con operazione di ingresso \square , operazione d'uscita \bigcirc e trasformazione del sistema $H[]$

Se vogliamo usare la teoria degli spazi vettoriali lineari in questo modo, le operazioni di ingresso e di uscita devono soddisfare, rispettivamente, i postulati algebrici della somma vettoriale e della moltiplicazione scalare. Questo significa, ad esempio, che le operazioni \square e \circ devono essere sia commutative che associative, cioè deve valere

$$\begin{aligned} x_1(n) \square x_2(n) &= x_2(n) \square x_1(n) \\ y_1(n) \circ y_2(n) &= y_2(n) \circ y_1(n) \end{aligned} \quad (10.3)$$

e

$$\begin{aligned} x_1(n) \square [x_2(n) \square x_3(n)] &= [x_1(n) \square x_2(n)] \square x_3(n) \\ y_1(n) \circ [y_2(n) \circ y_3(n)] &= [y_1(n) \circ y_2(n)] \circ y_3(n) \end{aligned} \quad (10.4)$$

Vi sono molte considerazioni matematiche di questo tipo da tenere presenti nel definire in modo appropriato gli spazi vettoriali e le trasformazioni [1,2]. Noi non ci addentreremo in una discussione rigorosa dei dettagli matematici ma descriveremo semplicemente il risultato fondamentale dell'applicazione della teoria degli spazi vettoriali a sistemi che obbediscono a un principio di sovrapposizione generalizzato.

Si può dimostrare che se gli ingressi del sistema costituiscono uno spazio vettoriale, con \square e $:$ che corrispondono alla somma vettoriale e alla moltiplicazione scalare, e se le uscite del sistema costituiscono uno spazio vettoriale, con \circ e \mathbf{L} che corrispondono alla somma vettoriale e alla moltiplicazione scalare, allora tutti i sistemi di questa classe possono essere rappresentati come una cascata di tre sistemi, come mostrato in fig. 10.2.

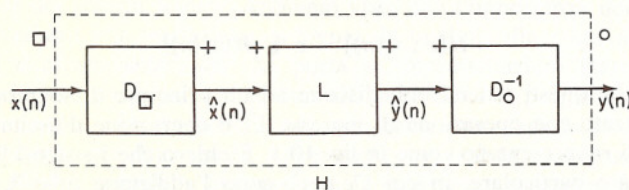


Fig. 10.2 Rappresentazione canonica di sistemi omomorfi.

Questa cascata di fig. 10.2 viene chiamata *rappresentazione canonica di sistemi omomorfi*. Il primo sistema, D_{\square} , ha la proprietà che

$$\begin{aligned} D_{\square}[x_1(n) \square x_2(n)] &= D_{\square}[x_1(n)] + D_{\square}[x_2(n)] \\ &= \hat{x}_1(n) + \hat{x}_2(n) \end{aligned} \quad (10.5a)$$

$$D_{\square}[c : x_1(n)] = c D_{\square}[x_1(n)] = c \hat{x}_1(n) \quad (10.5b)$$

Osserviamo che D_{\square} obbedisce a un principio di sovrapposizione generalizzato dove l'operazione d'ingresso è \square e l'operazione d'uscita è $+$. L'effetto del sistema D_{\square} è di trasformare la combinazione dei segnali $x_1(n)$ e $x_2(n)$ secondo la regola \square in una combinazione lineare convenzionale dei

segnali corrispondenti $D_{\square}[x_1(n)]$ e $D_{\square}[x_2(n)]$. Il sistema L è un sistema lineare convenzionale, per cui vale

$$\begin{aligned} L[\hat{x}_1(n) + \hat{x}_2(n)] &= L[\hat{x}_1(n)] + L[\hat{x}_2(n)] \\ &= \hat{y}_1(n) + \hat{y}_2(n) \end{aligned}$$

$$L[c \hat{x}_1(n)] = c L[\hat{x}_1(n)] = c \hat{y}_1(n)$$

Infine, il sistema D_{\circ}^{-1} esegue la trasformazione dall'addizione a \circ , in modo che risulta

$$\begin{aligned} D_{\circ}^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] &= D_{\circ}^{-1}[\hat{y}_1(n)] \circ D_{\circ}^{-1}[\hat{y}_2(n)] \\ &= y_1(n) \circ y_2(n) \end{aligned}$$

$$D_{\circ}^{-1}[c \hat{y}_1(n)] = c \mathbf{L} D_{\circ}^{-1}[\hat{y}_1(n)] = c \mathbf{L} y_1(n)$$

Poiché il sistema D_{\square} è fissato dalle operazioni \square e $:$, esso è caratteristico della classe ed è perciò chiamato il *sistema caratteristico* per l'operazione \square . Analogamente, D_{\circ} è il sistema caratteristico per l'operazione \circ . È chiaro, inoltre, che tutti i sistemi omomorfi con le stesse operazioni di ingresso e uscita differiscono solo nella parte lineare. Questo risultato è di fondamentale importanza, in quanto implica che, una volta determinati i sistemi caratteristici della classe, rimane solo da risolvere un problema di filtraggio lineare. Per esempio, se vogliamo riottenere $x_1(n)$ dal segnale

$$x(n) = x_1(n) \square x_2(n)$$

dobbiamo scegliere il sistema lineare in modo che la sua uscita $\hat{y}(n)$ sia

$$\hat{y}(n) = \hat{x}_1(n)$$

Allora, con $D_{\circ} = D_{\square}$, risulta

$$y(n) = D_{\square}^{-1}[\hat{x}_1(n)] = x_1(n)$$

In altri termini, per ottenere la perfetta separazione di $x_1(n)$ e $x_2(n)$, dobbiamo riuscire a separare perfettamente $\hat{x}_1(n)$ e $\hat{x}_2(n)$ usando un filtro lineare. Quanto bene si possa approssimare questa situazione ideale dipende dall'operazione \square e dalle proprietà dei segnali $x_1(n)$ e $x_2(n)$. Nel resto di questo capitolo limiteremo la trattazione a classi di sistemi per i quali le operazioni di ingresso e di uscita sono la stessa, ed in particolare alle due classi di sistemi omomorfi definiti scegliendo per questa operazione la moltiplicazione o la convoluzione. Esamineremo per ogni caso le rappresentazioni del sistema D_{\square} e le classi di segnali per i quali un'elaborazione di questo tipo appare vantaggiosa rispetto ad altre tecniche.

10.2 SISTEMI OMOMORFI MULTIPLICATIVI

Esistono molti problemi di trattamento dei segnali dove il segnale può essere rappresentato come il prodotto di due o più segnali componenti. Nella trasmissione di un segnale su un canale con evanescenza («fading»),

ad esempio, possiamo modellare questo effetto in termini di una componente a variazione lenta che moltiplica il segnale trasmesso. Per fare un altro esempio, un segnale modulato in ampiezza è rappresentato dal prodotto di un segnale portante e di una funzione involuppo, che vogliamo separare al ricevitore. Altri esempi riguardano la compressione di dinamica di segnali fonici e il trattamento di immagini, che discuteremo nel par. 10.3. In molti problemi di questo tipo un sistema lineare può essere del tutto inefficace per separare o per modificare in modo indipendente i segnali componenti. Al contrario, un sistema che obbedisca ad un principio di sovrapposizione generalizzato per la moltiplicazione può spesso essere impiegato con risultati sorprendenti. In questo paragrafo discutiamo la teoria di base dei sistemi omomorfi per la moltiplicazione e nel prossimo illustriamo le applicazioni di tali sistemi all'elaborazione numerica di im-

Consideriamo la classe di sistemi omomorfi che obbediscono ad un principio di sovrapposizione generalizzato in cui l'operazione \square è la moltiplicazione e l'operazione \circ è l'elevazione a potenza. In altri termini, ci occupiamo di segnali della forma

$$x(n) = [x_1(n)]^\alpha \cdot [x_2(n)]^\beta \quad (10.6)$$

Si verifica facilmente che queste sono scelte appropriate per la somma vettoriale e per la moltiplicazione scalare. Il sistema caratteristico per la moltiplicazione deve avere la proprietà

$$D \cdot [[x_1(n)]^\alpha \cdot [x_2(n)]^\beta] = \alpha D \cdot [x_1(n)] + \beta D \cdot [x_2(n)] \quad (10.7)$$

Una funzione che formalmente gode di questa proprietà è la funzione logaritmo. Ad esempio, se $x(n) = x_1(n) \cdot x_2(n)$, dove $x_1(n) > 0$ e $x_2(n) > 0$ per tutti i valori di n , allora risulta

$$\log [x_1(n) \cdot x_2(n)] = \log [x_1(n)] + \log [x_2(n)] \quad (10.8)$$

L'ingresso $x(n)$ può però non essere sempre positivo, e inoltre possiamo voler considerare segnali che sono complessi. Questi casi richiedono l'uso della funzione logaritmo complesso. Quindi la rappresentazione formale di sistemi canonici con la moltiplicazione come operazione di ingresso e di uscita è quella di fig. 10.3, dove le sequenze $x(n)$, $\hat{x}(n)$, $\hat{y}(n)$ e $y(n)$ sono in generale complesse.

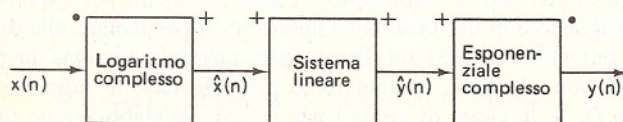


Fig. 10.3 Rappresentazione canonica per sistemi omomorfi con la moltiplicazione come operazione di ingresso e di uscita.

Consideriamo le proprietà della funzione logaritmo complesso. Se $x(n) = |x(n)|e^{j\arg[x(n)]}$ indica una sequenza complessa, allora il logaritmo complesso di $x(n)$ è definito come

$$\log [x(n)] = \log |x(n)| + j \arg [x(n)] \quad (10.9)$$

L'inverso di $\log[x(n)]$ è l'esponenziale complesso

$$e^{\log[x(n)]} = e^{\log|x(n)|} \cdot e^{j\arg[x(n)]} \quad (10.10)$$

È chiaro che si può aggiungere qualsiasi multiplo intero di 2π alla parte immaginaria del logaritmo complesso, cioè ad $\arg[x(n)]$, senza cambiare il risultato della (10.10). Perciò, se non si impongono ulteriori vincoli, il logaritmo complesso non è una trasformazione univoca. Poiché l'univocità è un requisito fondamentale nella definizione di un sistema, dobbiamo scegliere $\arg[x(n)]$ in modo da eliminare questa ambiguità. Inoltre, abbiamo il vincolo aggiuntivo che $\log[x(n)]$ sia definito in modo che valga la sovrapposizione generalizzata, cioè in modo che se $x(n) = x_1(n) \cdot x_2(n)$, risulti

$$\log [x(n)] = \log [x_1(n) \cdot x_2(n)] = \log [x_1(n)] + \log [x_2(n)]$$

Questo implica che sia

$$\log |x(n)| = \log |x_1(n)| + \log |x_2(n)| \quad (10.11)$$

e

$$\arg [x(n)] = \arg [x_1(n)] + \arg [x_2(n)] \quad (10.12)$$

Quindi l'ambiguità in $\arg[x(n)]$ deve essere eliminata in modo che sia soddisfatta la relazione (10.12).

Anche se di solito si elimina l'ambiguità nel logaritmo complesso sostituendo ad $\arg[x(n)]$ il suo valore principale, cioè il suo valore modulo 2π , non lo si può fare in questo caso perché la relazione (10.12) non sarebbe in generale soddisfatta. In altri termini, non è vero in generale che il valore principale della somma di due angoli è uguale alla somma dei loro rispettivi valori principali. Per risolvere l'ambiguità e soddisfare la (10.12) è necessario definire $\arg[x]$ in modo che sia una funzione continua di x . Questo approccio può essere sviluppato rigorosamente usando la teoria delle superfici di Riemann. In questo paragrafo non ci occuperemo dei dettagli di tale definizione del logaritmo complesso, in quanto i segnali delle classi di interesse nelle applicazioni che presenteremo sono non negativi, nel qual caso non esiste ambiguità poiché $\arg[x(n)]$ può sempre essere assunto nullo. Tuttavia, questo non sarà vero nel par. 10.4, dove esamineremo sistemi omomorfi per la convoluzione, ed in quel paragrafo discuteremo in dettaglio una definizione non ambigua del logaritmo complesso. Quindi per il momento assumiamo che sia disponibile una definizione non ambigua del logaritmo complesso per la realizzazione del sistema caratteristico relativo ai sistemi omomorfi moltiplicativi. Allora un particolare sistema omomorfo di questa classe differisce da tutti gli altri della medesima classe solo nella parte lineare. Se l'ingresso è del tipo (10.6), l'uscita del logaritmo complesso è

$$\hat{x}(n) = \alpha \hat{x}_1(n) + \beta \hat{x}_2(n) = \hat{x}_r(n) + j \hat{x}_i(n) \quad (10.13)$$

dove è

$$\hat{x}_1(n) = \log [x_1(n)] \quad \text{e} \quad \hat{x}_2(n) = \log [x_2(n)] \quad (10.14)$$

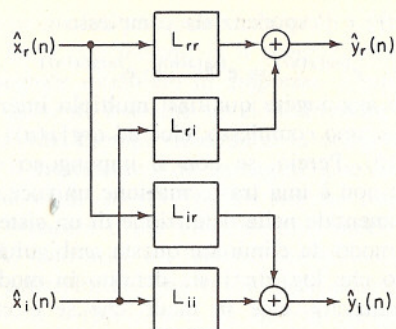


Fig. 10.4 Rappresentazione generale di un sistema lineare con ingresso e uscita complessi.

La forma più generale di un sistema lineare per l'elaborazione di ingressi complessi è mostrata in fig. 10.4, dove L_{rr} , L_{ri} , L_{ir} e L_{ii} indicano sistemi lineari reali. Se α e β sono numeri complessi, i sistemi lineari reali devono soddisfare le uguaglianze

$$L_{rr} = L_{ii} \quad \text{e} \quad L_{ri} = -L_{ir} \quad (10.15)$$

Se α e β sono reali, non ci sono vincoli.

Nelle applicazioni della teoria dei sistemi omomorfi moltiplicativi, occorre scegliere opportunamente il sistema lineare e la scelta dipende ovviamente dalla natura del segnale d'ingresso. La possibilità o meno di elaborare in modo utile segnali della forma

$$x(n) = x_1(n) \cdot x_2(n)$$

dipende dalla natura di $\hat{x}_1(n)$ e $\hat{x}_2(n)$, le componenti dell'uscita del sistema caratteristico. Ad esempio, se vogliamo separare le due componenti o elaborarle in maniera indipendente una dall'altra, allora i loro spettri in frequenza non devono avere sovrapposizioni significative. Questo implica che il trattamento omomorfo di segnali combinati tramite la moltiplicazione può essere utile tutte le volte che una componente è rapidamente variabile e l'altra varia lentamente. Questo è proprio quanto avviene nei casi della compressione di segnali fonici e del trattamento di immagini, dove il filtraggio omomorfo è stato applicato con successo [3], [4], [5], [6]. Nel prossimo paragrafo considereremo in dettaglio quest'ultimo esempio.

10.3 ELABORAZIONE OMOMORFA DI IMMAGINI

Un esempio dei vantaggi dell'applicazione dell'elaborazione omomorfa dei segnali si ha nel campo del miglioramento di qualità delle immagini. Quest'applicazione si basa su un modello delle immagini come prodotto di due componenti di base. Queste componenti possono essere modificate separatamente mediante il filtraggio omomorfo per ottenere contemporaneamente un miglioramento del contrasto e una compressione del campo dinamico.

10.3.1 Modello per la formazione delle immagini [6]

Le immagini sono formate dalla riflessione della luce, cioè si forma un'immagine quando l'energia luminosa proveniente da una sorgente di illuminazione viene riflessa da oggetti fisici. Quindi, il processo di formazione delle immagini può essere modellato come un processo moltiplicativo in cui una funzione illuminazione è moltiplicata per una funzione riflessione. Se le configurazioni bidimensionali di illuminazione e di riflessione sono rappresentate rispettivamente dalle funzioni $f_i(u, v)$ e $f_r(u, v)$ ¹, dove u e v sono variabili spaziali continue, allora l'immagine è espressa come

$$f(u, v) = f_i(u, v) \cdot f_r(u, v) \quad (10.16)$$

La componente di riflessione è sempre positiva ed è inoltre vincolata, per considerazioni fisiche, ad essere minore di uno:

$$0 < f(u, v) < f_i(u, v) < \infty \quad (10.17)$$

Poiché poi sia la componente di illuminazione che l'immagine corrispondono a distribuzioni di energia luminosa, risulta

$$0 < f_r(u, v) < 1 \quad (10.18)$$

Riassumendo, le immagini possono essere rappresentate come il prodotto di due componenti e, inoltre, le singole componenti sono sempre positive. Pertanto la struttura delle immagini sembra essere particolarmente adatta per l'elaborazione con un sistema omomorfo moltiplicativo.

10.3.2 Elaborazione numerica delle immagini

Le immagini vengono rappresentate nel discreto mediante una sequenza bidimensionale $x(m, n)$ ottenuta con un campionamento periodico:

$$x(m, n) = f(m \Delta u, n \Delta v)$$

dove Δu e Δv sono scelti in modo che non si verifichi in quantità significativa il fenomeno dell'*aliasing*. In base alla discussione precedente, $x(m, n)$ ha la rappresentazione

$$x(m, n) = x_i(m, n) \cdot x_r(m, n) \quad (10.19)$$

dove $x_i(m, n)$ e $x_r(m, n)$ sono le funzioni campionate rispettivamente di illuminazione e di riflettanza.

Il sistema omomorfo discreto per l'elaborazione delle immagini ha la forma canonica illustrata in fig. 10.5 Poiché è $x(m, n) > 0$, non occorre preoccuparsi dell'ambiguità del logaritmo complesso. Si può scrivere quindi

$$\begin{aligned} \hat{x}(m, n) &= \log [x(m, n)] = \log [x_i(m, n) \cdot x_r(m, n)] \\ &= \log x_i(m, n) + \log x_r(m, n) \\ &= \hat{x}_i(m, n) + \hat{x}_r(m, n) \end{aligned} \quad (10.20)$$

Stockam [6] ha osservato che l'uscita del sistema caratteristico ha una

¹ Nel par. 10.3 gli indici r ed i si riferiscono rispettivamente alla riflessione e all'illuminazione e non alle parti reale e immaginaria.

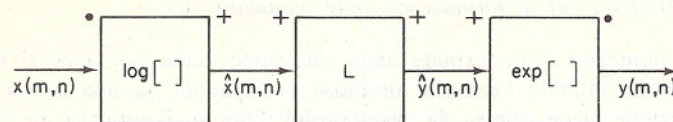


Fig. 10.5 Forma canonica di un sistema omomorfo per l'elaborazione delle immagini.

stretta corrispondenza con la configurazione di densità che si ha rappresentando un'immagine mediante diapositiva. Ha quindi senso chiamare $x(m, n)$ *rappresentazione intensità* e $\hat{x}(m, n)$ *rappresentazione densità* di un'immagine. Analogamente, $\hat{y}(m, n)$ e $y(m, n)$ possono essere chiamate *rappresentazioni densità e intensità dopo l'elaborazione*. È chiaro dalle proprietà dei sistemi lineari che risulta

$$\hat{y}(m, n) = \hat{y}_i(m, n) + \hat{y}_r(m, n) \quad (10.21)$$

dove $\hat{y}_i(m, n)$ e $\hat{y}_r(m, n)$ sono rispettivamente le densità d'illuminazione e di riflessione dopo l'elaborazione. Perciò l'intensità in uscita dopo l'elaborazione è

$$\begin{aligned} y(m, n) &= \exp [\hat{y}_i(m, n) + \hat{y}_r(m, n)] \\ &= \exp [\hat{y}_i(m, n)] \cdot \exp [\hat{y}_r(m, n)] \\ &= y_i(m, n) y_r(m, n) \end{aligned} \quad (10.22)$$

Poiché $\hat{y}_i(m, n)$ e $\hat{y}_r(m, n)$ sono entrambe reali (assumendo che la risposta all'impulso del sistema lineare sia reale), $y_i(m, n)$, $y_r(m, n)$ e $y(m, n)$ sono tutte positive. Quindi risulta soddisfatto il vincolo che in un'immagine fisica le intensità devono essere strettamente positive. Questo comportamento è garantito per tutti i sistemi che hanno la forma di fig. 10.5. Ciò non avviene quando si elabora con un sistema lineare la rappresentazione intensità di un'immagine: in questo caso non è infatti garantito che le uscite siano positive.

Avendo dimostrato che i sistemi omomorfi moltiplicativi si adattano molto bene alla struttura delle immagini, resta il problema di come scegliere il sistema lineare di fig. 10.5. La scelta dipende, naturalmente, dalle proprietà delle componenti $\hat{x}_i(m, n)$ e $\hat{x}_r(m, n)$ della densità. Fortunatamente, le componenti di illuminazione e di riflessione hanno caratteristiche nettamente diverse. Infatti, in generale, l'illuminazione non varia rapidamente muovendosi lungo le varie parti di una scena, anche se le ombre corrispondono, ovviamente, a variazioni brusche dell'illuminazione. D'altro canto, se la scena contiene sufficiente dettaglio, la componente di riflessione varierà rapidamente a causa dei bordi netti e dei cambiamenti di trama e dimensione. È quindi ragionevole assumere che l'illuminazione sia un segnale a bassa frequenza (spaziale) e invece la componente di riflessione sia un segnale ad alta frequenza. Di conseguenza, la densità d'illuminazione avrà variazioni lente, mentre la densità di riflessione varierà con rapidità. I grafici della trasformata di Fourier di un'immagine o del suo logaritmo sono caratterizzati tipicamente da un picco a bassa frequenza e

da un livello con piccole variazioni alle alte frequenze: ciò sembra confermare implicitamente la validità delle assunzioni fatte prima, anche se associare il picco a bassa frequenza unicamente all'illuminazione e le alte frequenze unicamente alla riflessione sarebbe una semplificazione eccessiva.

Anche se lente, le variazioni di illuminazione possono però essere molto notevoli nell'ambito di una scena, producendo quindi un vasto campo dinamico nell'immagine. Questo fatto pone problemi non indifferenti nella trasmissione di immagini su un canale di comunicazione o nella loro registrazione su un mezzo quale la pellicola fotografica. Di qui l'importanza di elaborare la componente di illuminazione per ridurre il campo dinamico.

Un altro problema importante nel trattamento delle immagini è il miglioramento del contrasto, cioè il filtraggio dell'immagine in modo che risultino più netti i contorni degli oggetti della scena. Poiché questi contorni si manifestano essenzialmente come brusche variazioni nella funzione di riflessione, occorre operare proprio su quest'ultima se vogliamo alterare il contrasto di un'immagine.

Per vedere come si può ridurre il campo dinamico e migliorare il contrasto usando il sistema canonico di fig. 10.5, assumiamo dapprima che il sistema lineare sia un guadagno ideale γ indipendente dalla frequenza, per cui è

$$\hat{y}(m, n) = \gamma \hat{x}(m, n)$$

L'immagine d'uscita sarà allora

$$y(m, n) = [x(m, n)]^\gamma = [x_i(m, n)]^\gamma \cdot [x_r(m, n)]^\gamma \quad (10.23)$$

Il campo dinamico dell'immagine può essere ridotto diminuendo la variazione della componente di illuminazione e questo implica chiaramente una scelta di $\gamma < 1$. Per aumentare invece il contrasto, occorre elaborare la funzione riflessione in modo che cresca il rapporto tra due intensità date, e questo implica un $\gamma > 1$.

La trasformazione rappresentata dalla (10.23), che corrisponde ad un semplice guadagno per il sistema lineare, può essere ottenuta fotograficamente; tuttavia questa scelta per il sistema lineare è spesso una soluzione insoddisfacente in quanto si è visto che per $\gamma < 1$ si riduce il contrasto insieme al campo dinamico, mentre per $\gamma > 1$ insieme al contrasto aumenta anche il campo dinamico. Se però ricordiamo che l'illuminazione è un segnale a bassa frequenza e la riflessione è un segnale ad alta frequenza, è chiaro che un filtro lineare invariante alla traslazione che applichi guadagni diversi alle alte e alle basse frequenze rappresenta una scelta migliore. Ad esempio, la fig. 10.6 riporta una sezione radiale di una risposta in frequenza a simmetria circolare. Questo sistema offre la possibilità, almeno approssimativamente, di migliorare il contrasto e nello stesso tempo ridurre il campo dinamico; in altri termini, $y(m, n)$ è, in prima approssimazione, della forma

$$y(m, n) = [x_i(m, n)]^{\gamma_i} [x_r(m, n)]^{\gamma_r} \quad (10.24)$$

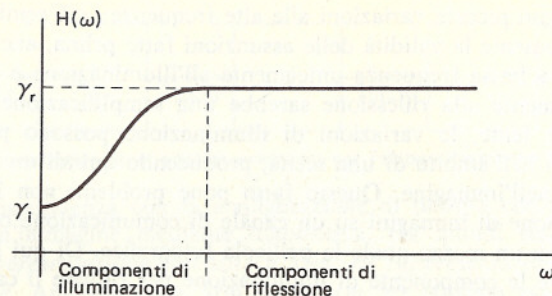


Fig. 10.6 Sezione di una risposta in frequenza a simmetria circolare da usare per la parte lineare di un filtro omomorfo per immagini, allo scopo di ottenere contemporaneamente miglioramento di contrasto e compressione del campo dinamico.

Quindi per ottenere contemporaneamente una riduzione del campo dinamico e un miglioramento del contrasto, γ_i va scelto minore dell'unità e γ_r va scelto maggiore dell'unità.

Nelle fig. 10.7 e 10.8 sono presentati due esempi di immagini filtrate in questo modo per avere contemporaneamente la compressione di dinamica e il miglioramento del contrasto. Il guadagno a bassa frequenza del filtro, γ_i , è stato scelto pari a 0.5, e il guadagno ad alta frequenza, γ_r , pari a 2, il che corrisponde a porre $\gamma_i = 0.5$ e $\gamma_r = 2$ nella relazione (10.24). In fig. 10.7 sono rappresentate le scene originali e nella 10.8 le stesse dopo l'elaborazione. Altri esempi del tipo di filtraggio discusso qui si trovano in [3].

10.4 SISTEMI OMOMORFI PER LA CONVOLUZIONE

Vi sono moltissimi problemi di elaborazione dei segnali in cui i segnali sono combinati tramite l'operazione di convoluzione. Quando, ad esempio, si effettua una trasmissione o una registrazione in condizioni che danno luogo a cammini multipli o riverberi, l'effetto della distorsione che ne consegue può essere modellato in termini di una convoluzione tra rumore e segnale desiderato. Un altro caso è quello dell'elaborazione della voce, dove spesso interessa separare gli effetti della risposta all'impulso del tratto vocale e dell'eccitazione, dalla cui convoluzione si può considerare generata la forma d'onda della voce, almeno in un'analisi a tempo breve. Altri esempi sono la separazione di funzioni densità di probabilità che risultano convolute come conseguenza della somma di processi casuali indipendenti, o l'elaborazione di segnali sismici ottenuti quando un'esplosione genera un impulso di energia sismica che si propaga nella terra.

Una tecnica comunemente usata per separare le componenti di questi segnali, cioè per eseguire la deconvoluzione, è il filtraggio lineare inverso. Sfortunatamente, essendo i sistemi lineari non « adattati » alla struttura delle combinazioni mediante convoluzione, il filtraggio inverso richiede

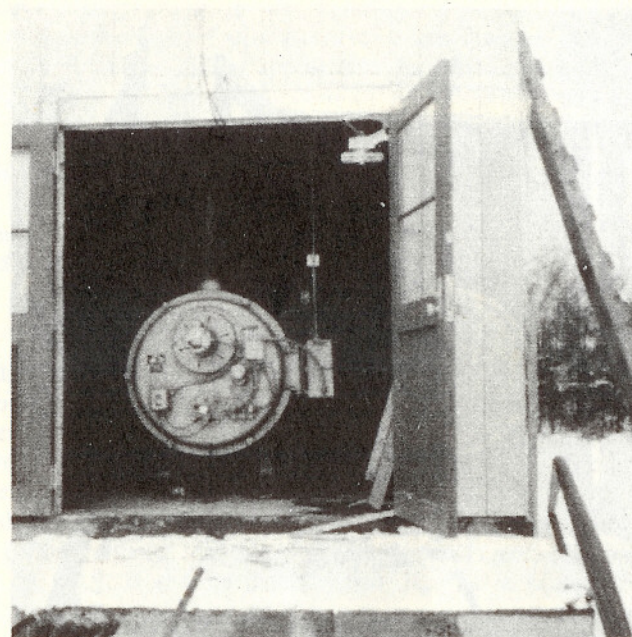


Fig. 10.7 Due immagini originali.

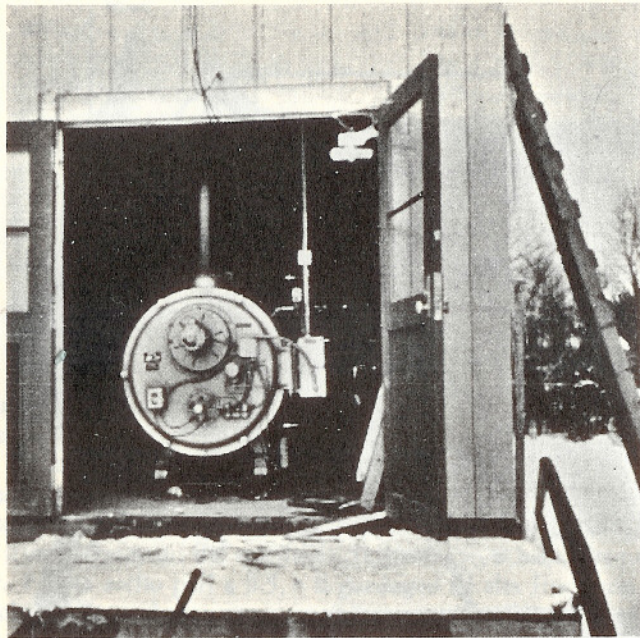


Fig. 10.8 Immagini di fig. 10.7 dopo il filtraggio per ottenere contemporaneamente compressione di dinamica e miglioramento del contrasto.

la conoscenza dettagliata di uno dei segnali componenti. Come alternativa, siamo portati a considerare una classe di sistemi omomorfi che obbediscono ad un principio di sovrapposizione generalizzato per la convoluzione [4, 7, 8].

10.4.1 Sistema canonico

Consideriamo la combinazione di sequenze per mezzo della convoluzione discreta, cioè

$$x(n) = \sum_{k=-\infty}^{\infty} x_1(k)x_2(n-k) = x_1(n) * x_2(n) \quad (10.25)$$

Si dimostra facilmente che la convoluzione discreta soddisfa i postulati algebrici per la somma vettoriale e di conseguenza è una scelta ammissibile per una classe di sistemi omomorfi. La moltiplicazione scalare per un intero a corrisponde alla convoluzione ripetuta di $x(n)$ con sé stessa a volte, e la moltiplicazione scalare quando a non è intero è una generalizzazione di questa operazione [1, 8].

La forma canonica dei filtri omomorfi per la convoluzione è illustrata in fig. 10.9. Il sistema caratteristico, D_* , ha le proprietà seguenti

$$D_*[x_1(n) * x_2(n)] = D_*[x_1(n)] + D_*[x_2(n)] = \hat{x}_1(n) + \hat{x}_2(n) \quad (10.26a)$$

$$D_*[c : x_1(n)] = cD_*[x_1(n)] = c\hat{x}_1(n) \quad (10.26b)$$

Il sistema L è un sistema lineare e D_*^{-1} è l'inverso di D_* . Quindi, se riusciamo a determinare il sistema D_* , abbiamo una rappresentazione di tutti i sistemi che soddisfano un principio di sovrapposizione generalizzato per la convoluzione. Nel resto di questo paragrafo e nei par. 10.5 e 10.6 studieremo in dettaglio le proprietà del sistema D_* e le proprietà dei segnali $\hat{x}(n)$. Nel par. 10.7 utilizzeremo poi queste proprietà in alcune applicazioni della deconvoluzione omomorfa.

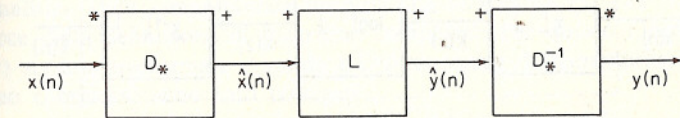


Fig. 10.9 Forma canonica per i filtri omomorfi con la convoluzione come operazione di ingresso e di uscita.

10.4.2 Rappresentazioni matematiche del sistema caratteristico D_*

Il punto chiave per la rappresentazione matematica del sistema caratteristico D_* è il fatto che la trasformata z della relazione (10.25) è

$$X(z) = X_1(z) \cdot X_2(z) \quad (10.27)$$

In altri termini, l'operazione di trasformata z $\zeta[x(n)]$ può essere vista come una trasformazione omomorfa avente la convoluzione come operazione d'ingresso e la moltiplicazione come operazione d'uscita, come illustrato in fig. 10.10. Usando la trasformazione z , le combinazioni me-

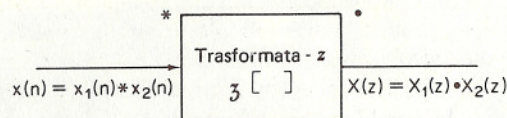


Fig. 10.10 Trasformata z rappresentata come una trasformazione omomorfa dalla convoluzione alla moltiplicazione.

diate convoluzione possono essere trasformate in combinazioni mediante moltiplicazione, che possono allora essere filtrate con un sistema omomorfo moltiplicativo. Perciò, se rappresentiamo i segnali con le loro trasformate z , il sistema canonico di fig. 10.9 può essere sostituito da quello di fig. 10.11. Poiché normalmente la funzione $X(z)$ è complessa, occorre usare qui il logaritmo complesso.

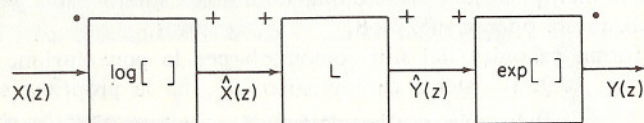


Fig. 10.11 Sistema di fig. 10.9 con i segnali rappresentati in termini delle loro trasformate z .

Quando i segnali sono rappresentati come sequenze invece che tramite le loro trasformate z , si può rappresentare formalmente il sistema caratteristico D_* come in fig. 10.12, dove ζ^{-1} è la trasformata z inversa. È interessante notare che sia ζ che ζ^{-1} sono trasformazioni lineari in senso convenzionale e nello stesso tempo trasformazioni omomorfe tra spazi vettoriali caratterizzati dalla convoluzione e dalla moltiplicazione.

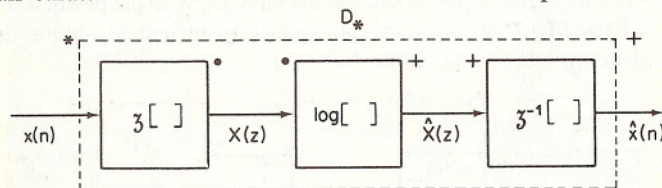


Fig. 10.12 Rappresentazione del sistema caratteristico D_* .

Vi sono alcune ipotesi importanti implicite nella rappresentazione del sistema caratteristico D_* come in fig. 10.12. Innanzitutto, il logaritmo complesso $\log[X(z)]$ deve essere univocamente definito, in modo che se è

$$X(z) = X_1(z) \cdot X_2(z)$$

risulti

$$\begin{aligned} \hat{X}(z) &= \log [X_1(z) \cdot X_2(z)] \\ &= \log [X_1(z)] + \log [X_2(z)] \end{aligned}$$

Inoltre, $\hat{X}(z)$ deve essere una trasformata z valida. In terzo luogo, affinché $\hat{x}(n)$ sia definito univocamente, occorre scegliere una regione di convergenza per $\hat{X}(z) = \log[X(z)]$. Consideriamo per prima la scelta della regione di convergenza. Assumiamo che $x(n)$ e $\hat{x}(n)$ siano entrambe sequenze reali stabili. Questa ipotesi, che è chiaramente molto ragionevole da un punto di vista pratico, non è in realtà restrittiva poiché bastano lievi modifiche per includere il caso di $x(n)$ o $\hat{x}(n)$ instabili. Le regioni di convergenza sia di $X(z)$ che di $\hat{X}(z)$ devono quindi includere il circolo unitario.

Se $\hat{X}(z) = \log[X(z)]$ è una trasformata z , allora deve avere un'espansione in serie di Laurent,

$$\hat{X}(z) = \log[X(z)] = \sum_{n=-\infty}^{\infty} \hat{x}(n)z^{-n}$$

con una regione di convergenza che racchiude il circolo unitario. Vale a dire che $\hat{X}(z)$ deve essere analitica in una regione che comprende il circolo unitario. Esprimiamo $\hat{X}(z)$ su quest'ultimo come

$$\hat{X}(e^{j\omega}) = \hat{X}_R(e^{j\omega}) + j\hat{X}_I(e^{j\omega})$$

Poiché $\hat{x}(n)$ è reale, $\hat{X}_R(e^{j\omega})$ deve essere una funzione pari di ω , e $\hat{X}_I(e^{j\omega})$ deve essere una funzione dispari di ω . Inoltre, $\hat{X}(e^{j\omega})$ deve essere una funzione periodica di ω con periodo 2π . La conseguenza più importante dell'analiticità di $\hat{X}(z)$ sul circolo unitario è che $\hat{X}(e^{j\omega})$ deve essere una funzione continua di ω . Dal fatto che è

$$\hat{X}(e^{j\omega}) = \log |X(e^{j\omega})| + j \arg [X(e^{j\omega})]$$

segue che

$$\hat{X}_R(e^{j\omega}) = \log |X(e^{j\omega})|$$

e

$$\hat{X}_I(e^{j\omega}) = \arg [X(e^{j\omega})]$$

devono essere funzioni continue di ω . Purché $X(z)$ non abbia zeri sul circolo unitario, la continuità di $\hat{X}_R(e^{j\omega})$ è assicurata dal fatto che $X(e^{j\omega})$ è analitica sulla circonferenza unitaria. La continuità di $\hat{X}_I(e^{j\omega})$ dipende invece dalla definizione di logaritmo complesso. Per questo l'imporre che $\hat{X}(z)$ sia una trasformata z valida e l'eliminazione dell'ambiguità del logaritmo complesso sono fatti collegati.

Il problema dell'univocità e dell'analiticità del logaritmo complesso è illustrato dalla fig. 10.13. In fig. 10.13(a) si vede una tipica curva di fase per la trasformata z , $X(z)$, valutata sul circolo unitario. Se $X(z)$ è il prodotto di due trasformate z , allora si può pensare questa curva come la somma delle curve continue di fase delle trasformate componenti. La fig. 10.13(b) mostra il valore principale della fase di $X(z)$. Entrambe le curve sono rappresentazioni della fase di $X(z)$ in quanto è

$$j \arg X(z) = j \text{ARG} X(z)$$

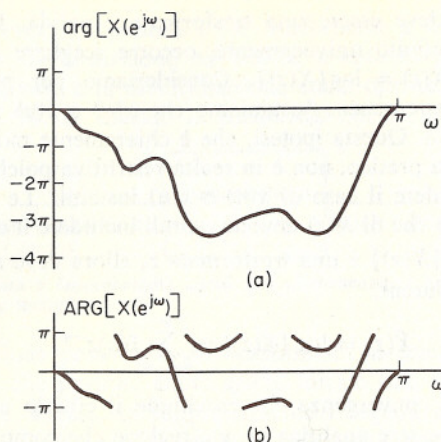


Fig. 10.13 (a) tipica curva di fase per una trasformata z valutata sul circolo unitario; (b) valore principale della curva di fase di (a).

Si vede però facilmente che la curva del valore principale potrebbe non corrispondere in generale alla somma dei valori principali delle fasi delle trasformate componenti, e per di più $\text{ARG}[X(z)]$ è discontinuo e quindi non può soddisfare il vincolo di continuità che deriva dal dover essere $\hat{X}(z)$ analitica sul circolo unitario.

Un modo di affrontare i problemi presentati dal logaritmo complesso è attraverso il concetto di superficie di Riemann. Un'altra via per definire il logaritmo complesso è assumere che il logaritmo complesso continuo si ottenga per integrazione della sua derivata. Nell'ipotesi di logaritmo complesso a un sol valore e differenziabile, possiamo scrivere

$$\frac{d}{dz} \log [X(z)] = \frac{1}{X(z)} \frac{dX(z)}{dz} = \frac{d\hat{X}(z)}{dz} \quad (10.28)$$

Valutando questa derivata logaritmica sul circolo unitario si ottiene

$$\hat{X}'(e^{j\omega}) = \frac{X'(e^{j\omega})}{X(e^{j\omega})} = \hat{X}'_R(e^{j\omega}) + j\hat{X}'_I(e^{j\omega})$$

dove l'apice indica la derivata rispetto ad ω . Da questa espressione si ha

$$\frac{d\hat{X}_I(e^{j\omega})}{d\omega} = \frac{X_R(e^{j\omega})\hat{X}'_I(e^{j\omega}) - X_I(e^{j\omega})\hat{X}'_R(e^{j\omega})}{X_R^2(e^{j\omega}) + X_I^2(e^{j\omega})} \quad (10.29)$$

Integrando la (10.29) rispetto ad ω possiamo applicare la condizione

$$\arg [X(e^{j\omega})]_{\omega=0} = 0$$

per imporre che $\arg[X(e^{j\omega})]$ sia una funzione continua e dispari di ω .

La definizione del logaritmo complesso è stata discussa accuratamente per due ragioni. La prima è che è importante capire a fondo il logaritmo complesso in modo che le manipolazioni formali che faremo nel seguito

di questo paragrafo e in paragrafi successivi possano essere fatte con la certezza della loro validità. In secondo luogo, i problemi di ambiguità nella definizione del logaritmo complesso corrispondono a importanti problemi di calcolo. Questi ultimi saranno esaminati in un successivo paragrafo sulla realizzazione di sistemi omomorfi per la convoluzione.

La rappresentazione matematica di fig. 10.12 è stata ricavata in base al fatto che la teoria dei sistemi omomorfi moltiplicativi del par. 10.2 può essere applicata alle trasformate z di una convoluzione. Partendo da questa rappresentazione e dall'implicita assunzione dell'analiticità di $\log[X(z)]$, possiamo ottenere altre due rappresentazioni del sistema D_* usando la derivata logaritmica.

Assumendo che $\log[X(z)]$ sia analitico, risulta

$$\hat{X}'(z) = \frac{X'(z)}{X(z)} \quad (10.30)$$

dove ' indica la derivazione rispetto a z . Si dimostra facilmente che è

$$z\hat{X}'(z) = \sum_{n=-\infty}^{\infty} [-n\hat{x}(n)]z^{-n} = \frac{zX'(z)}{X(z)} \quad (10.31)$$

da cui segue

$$-n\hat{x}(n) = \frac{1}{2\pi j} \oint_C \frac{zX'(z)}{X(z)} z^{n-1} dz$$

dove C è un percorso chiuso nella regione di convergenza di $\hat{X}(z)$. Esprimendo $\hat{x}(n)$ si ottiene

$$\hat{x}(n) = \frac{-1}{2\pi j n} \oint_C \frac{zX'(z)}{X(z)} z^{n-1} dz, \quad n \neq 0 \quad (10.32)$$

Il valore di $\hat{x}(0)$ si ricava notando che è

$$\begin{aligned} \hat{x}(0) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_R(e^{j\omega}) d\omega + \frac{j}{2\pi} \int_{-\pi}^{\pi} \hat{X}_I(e^{j\omega}) d\omega \end{aligned}$$

Poiché $\hat{X}_I(e^{j\omega})$ è una funzione dispari di ω , questa espressione diventa

$$\begin{aligned} \hat{x}(0) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_R(e^{j\omega}) d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(e^{j\omega})| d\omega \end{aligned} \quad (10.33)$$

Partendo dalla relazione (10.31) possiamo anche derivare un'equazione alle differenze che rappresenta il sistema D_* . Riordinando la (10.31) si ha

$$zX'(z) = z\hat{X}'(z) \cdot X(z)$$

La trasformata z inversa di questa espressione è

$$nx(n) = \sum_{k=-\infty}^{\infty} k\hat{x}(k)x(n-k) \quad (10.34)$$

e dividendo per n si ottiene

$$x(n) = \sum_{k=-\infty}^{\infty} \left(\frac{k}{n}\right) \hat{x}(k)x(n-k), \quad n \neq 0 \quad (10.35)$$

Abbiamo quindi ricavato una relazione implicita tra $\hat{x}(n)$ e $x(n)$. Sotto certe condizioni questa espressione può essere trasformata in una formula ricorsiva che si può usare per il calcolo. Formule di questo tipo sono discusse nel par. 10.5.

10.4.3 Sistema caratteristico inverso D_*^{-1}

La rappresentazione matematica del sistema caratteristico inverso D_*^{-1} segue in modo semplice dalla rappresentazione di D_* ed è illustrato in fig. 10.14. È per definizione

$$D_*^{-1}[D_*[x(n)]] = x(n)$$

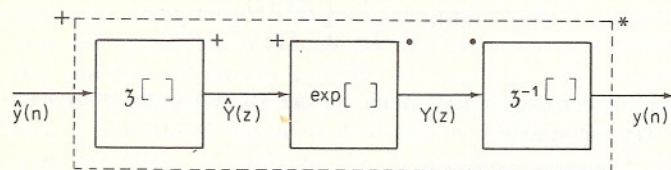


Fig. 10.14 Rappresentazione del sistema D_*^{-1} .

e perciò sia $\hat{y}(n)$ che $y(n)$ devono essere sequenze stabili, in quanto $x(n)$ e $\hat{x}(n)$ sono per ipotesi entrambe stabili. Le regioni di convergenza di $Y(z)$ e di $\hat{Y}(z)$ devono allora comprendere il cerchio unitario, per cui si ha

$$y(n) = \frac{1}{2\pi j} \oint_{C'} Y(z) z^{n-1} dz$$

dove C' è il cerchio unitario e

$$Y(z) = \exp[\hat{Y}(z)]$$

Per fortuna la funzione esponenziale complesso non presenta problemi di unicità e se $\hat{Y}(z)$ è analitica sul cerchio unitario lo è anche $\exp[\hat{Y}(z)]$.

10.4.4 Il sistema lineare L

Avendo fornito una rappresentazione matematica del sistema caratteristico D_* e del suo inverso, nel sistema canonico di fig. 10.9 resta solo da specificare il sistema L . In linea teorica, nel sistema canonico di fig. 10.9

può essere usato qualsiasi sistema che soddisfa il principio di sovrapposizione rispetto all'addizione. Sul piano pratico, tuttavia, si è trovato che è veramente utile soltanto una particolare classe di sistemi lineari. Ricordiamo che nel caso dei sistemi moltiplicativi la scelta opportuna per il sistema lineare è stata quella di un sistema lineare tempo-invariante. Ricordiamo anche dalla fig. 10.11 che se i segnali sono rappresentati dalle loro trasformate z , allora si può pensare ad un sistema lineare che agisce sul logaritmo complesso della trasformata z . In altre parole, la classe dei sistemi omomorfi per la convoluzione è analoga alla classe dei sistemi omomorfi per la moltiplicazione; i ruoli del dominio del tempo e del dominio della frequenza sono però, in un certo senso, scambiati. Quindi, anche se teoricamente si possono usare sistemi lineari tempo-invarianti, è particolarmente interessante prendere in esame la classe dei sistemi lineari frequenza-invarianti, per i quali risulta

$$\hat{Y}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\theta}) L(e^{j(\omega-\theta)}) d\theta \quad (10.36)$$

Per un sistema di questo tipo la trasformata di Fourier dell'uscita si ottiene dal logaritmo complesso $\hat{X}(e^{j\omega}) = \log[X(e^{j\omega})]$ per mezzo di una convoluzione periodica con variabile continua. In alternativa, tale sistema ha una rappresentazione nel dominio del tempo come

$$\hat{y}(n) = l(n)\hat{x}(n) \quad (10.37)$$

dove $l(n)$ è la trasformata inversa di $L(e^{j\omega})$. Poiché si assume che $x(n)$, $\hat{x}(n)$, $\hat{y}(n)$ e $y(n)$ siano tutte sequenze reali e stabili, ne segue che $l(n)$ deve essere reale, e in generale deve anche essere stabile. Ciò implica che $L(z)$, la trasformata z di $l(n)$, ha una regione di convergenza che comprende il cerchio unitario e che le parti reale e immaginaria di $L(e^{j\omega})$ sono funzioni rispettivamente pari e dispari di ω .

Si potrebbe naturalmente approfondire di più il problema del perché questa particolare classe di sistemi lineari sia così utile. Tuttavia, questo problema e quelli relativi al progetto di tali sistemi trovano soluzione solo alla luce di ulteriori discussioni sulle proprietà di $\hat{x}(n)$ o di $\log[X(e^{j\omega})]$. Questo è l'argomento del par. 10.5.

10.4.5 Nota sulla terminologia

Concludiamo questo paragrafo con una breve nota storica che aiuterà a definire la terminologia per il resto del capitolo. Nel 1963 Bogert, Healy e Tukey pubblicarono un lavoro dal titolo piuttosto insolito [9]. In questo articolo essi notarono che il logaritmo dello spettro di potenza di un segnale contenente un'eco ha una componente periodica additiva dovuta all'eco e quindi la trasformata di Fourier del logaritmo dello spettro di potenza deve avere un picco in corrispondenza del ritardo d'eco. Essi chiamarono questa funzione il *cepstrum*, anagrammando il termine inglese per spettro,

in quanto « In generale ci troviamo ad operare nel campo della frequenza con metodi tipici del dominio del tempo e viceversa » [9]. Bogert e gli altri proseguirono definendo un intero vocabolario relativo a questa nuova tecnica di elaborazione dei segnali; però ha avuto vasta diffusione solo il termine « cepstrum ». Siccome lo spettro di potenza è la trasformata di Fourier della funzione di autocorrelazione ed è sempre positivo, si può considerare il cepstrum come l'uscita del sistema caratteristico D_* quando l'ingresso è un'autocorrelazione. Essendo lo spettro di potenza sempre positivo, in questo caso basta calcolare il logaritmo reale. In generale dobbiamo usare il logaritmo complesso e trasformate di Fourier complesse: così, per sottolineare la relazione e mantenere la distinzione, chiamiamo l'uscita del sistema caratteristico il *cepstrum complesso*. Ci preme aggiungere che il *cepstrum complesso* $\hat{x}(n)$ è, ovviamente, reale per ingressi $x(n)$ reali. Riserveremo l'uso del termine « cepstrum » ai casi in cui si usa solo il logaritmo reale.

10.5 PROPRIETÀ DEL CEPSTRUM COMPLESSO

Proseguendo lo studio dei sistemi omomorfi per la convoluzione, esamineremo le proprietà del cepstrum complesso, ovvero di $\log[X(e^{j\omega})]$, nel caso di una particolare e interessante classe di segnali d'ingresso, quella delle sequenze esponenziali.

10.5.1 Sequenze esponenziali

La classe di sequenze aventi trasformata z razionale ha le fortunate proprietà di essere utile e di prestarsi ad un'analisi semplice. Consideriamo allora sequenze d'ingresso $x(n)$ le cui trasformate z sono della forma

$$X(z) = \frac{Az^r \prod_{k=1}^{m_i} (1 - a_k z^{-1}) \prod_{k=1}^{m_o} (1 - b_k z)}{\prod_{k=1}^{p_i} (1 - c_k z^{-1}) \prod_{k=1}^{p_o} (1 - d_k z)} \quad (10.38)$$

dove $|a_k|$, $|b_k|$, $|c_k|$ e $|d_k|$ sono tutti minori dell'unità, in modo che i fattori della forma $(1 - a_k z^{-1})$ e $(1 - c_k z^{-1})$ corrispondono a zeri e poli interni al cerchio unitario, e i fattori della forma $(1 - b_k z)$ e $(1 - d_k z)$ corrispondono a zeri e poli esterni al cerchio unitario. Queste trasformate z sono caratteristiche di sequenze costituite di una somma di sequenze esponenziali. Nel caso particolare in cui non vi siano poli, cioè il denominatore dell'espressione (10.38) sia unitario, la (10.38) corrisponde a una sequenza di lunghezza finita.

Se si calcola $\log[X(z)]$ come chiarito nel par. 10.4, si può scrivere formalmente

$$\hat{X}(z) = \log[A] + \log[z^r] + \sum_{k=1}^{m_i} \log(1 - a_k z^{-1}) + \sum_{k=1}^{m_o} \log(1 - b_k z) - \sum_{k=1}^{p_i} \log(1 - c_k z^{-1}) - \sum_{k=1}^{p_o} \log(1 - d_k z) \quad (10.39)$$

Se interpretiamo ogni termine della (10.39) come una trasformata z , le proprietà di $\hat{x}(n)$ dipenderanno dall'insieme delle proprietà delle trasformate inverse di ogni termine.

Per sequenze reali A è reale e se è anche positiva il primo termine, $\log[A]$, dà contributi solamente a $\hat{x}(0)$. Più precisamente, risulta (v. probl. 11 di questo capitolo)

$$\hat{x}(0) = \log|A| \quad (10.40)$$

Se A è negativa, è più difficile determinare il contributo del termine $\log[A]$ al cepstrum complesso. Analogamente, il termine z^r corrisponde solo a un ritardo o un anticipo della sequenza $x(n)$. Se $r = 0$, questo termine scompare dall'espressione (10.39). Se invece è $r \neq 0$, ci sarà un contributo non nullo al cepstrum complesso. Anche se i casi di A negativa e/o di $r \neq 0$ possono essere trattati formalmente [8], il farlo non sembra offrire vantaggi reali, perché di fronte al prodotto di due trasformate del tipo (10.38) non possiamo aspettarci di riuscire a determinare quanto sia il contributo di ciascuna componente ad A o ad r . Situazione analoga si ha nel filtraggio lineare quando siano stati sommati due segnali aventi ciascuno una componente continua. In pratica si evita quindi questo problema misurando il segno di A e il valore di r e cambiando poi l'ingresso in modo che la sua trasformata z sia della forma

$$X(z) = \frac{|A| \prod_{k=1}^{m_i} (1 - a_k z^{-1}) \prod_{k=1}^{m_o} (1 - b_k z)}{\prod_{k=1}^{p_i} (1 - c_k z^{-1}) \prod_{k=1}^{p_o} (1 - d_k z)} \quad (10.41)$$

Allo stesso modo la (10.39) diventa

$$\hat{X}(z) = \log|A| + \sum_{k=1}^{m_i} \log(1 - a_k z^{-1}) + \sum_{k=1}^{m_o} \log(1 - b_k z) - \sum_{k=1}^{p_i} \log(1 - c_k z^{-1}) - \sum_{k=1}^{p_o} \log(1 - d_k z) \quad (10.42)$$

Ad eccezione del termine $\log|A|$, che abbiamo già trattato, tutti i termini della (10.42) sono della forma $\log(1 - \alpha z^{-1})$ e $\log(1 - \beta z)$. Tenendo presente che questi fattori devono rappresentare trasformate z con regioni di

convergenza che comprendono il cerchio unitario, possiamo effettuare le espansioni in serie di potenze

$$\log(1 - \alpha z^{-1}) = - \sum_{n=1}^{\infty} \frac{\alpha^n}{n} z^{-n}, \quad |z| > |\alpha| \quad (10.43)$$

e

$$\log(1 - \beta z) = - \sum_{n=1}^{\infty} \frac{\beta^n}{n} z^n, \quad |z| < |\beta^{-1}| \quad (10.44)$$

Sulla base di queste espressioni è chiaro che, per ingressi con trasformate z razionali come la (10.41), $\hat{x}(n)$ ha la forma generale

$$\hat{x}(n) = \log |A| \quad n = 0 \quad (10.45a)$$

$$= - \sum_{k=1}^{m_i} \frac{a_k^n}{n} + \sum_{k=1}^{p_i} \frac{c_k^n}{n}, \quad n > 0 \quad (10.45b)$$

$$= \sum_{k=1}^{m_o} \frac{b_k^{-n}}{n} - \sum_{k=1}^{p_o} \frac{d_k^{-n}}{n}, \quad n < 0 \quad (10.45c)$$

Si noti che nel caso particolare di sequenze di durata finita mancherebbe il secondo termine sia nella (10.45b) che nella (10.45c). Dalle espressioni (10.45) ricaviamo le seguenti proprietà del cepstrum complesso.

P1. Il cepstrum complesso tende a zero con velocità pari almeno a $1/n$: più precisamente, risulta

$$|\hat{x}(n)| < C \left| \frac{\alpha^n}{n} \right|, \quad -\infty < n < \infty$$

dove C è una costante ed α è il massimo fra $|a_k|$, $|b_k|$, $|c_k|$ e $|d_k|$. Questa proprietà è anche evidente dalla (10.32).

P2. Se $x(n)$ è a fase minima (né poli né zeri esterni al cerchio unitario), si ha

$$\hat{x}(n) = 0, \quad n < 0$$

P3. Se $x(n)$ è a fase massima (né poli né zeri interni al cerchio unitario), si ha

$$\hat{x}(n) = 0, \quad n > 0$$

P4. Se $x(n)$ è di durata finita, $\hat{x}(n)$ avrà ciò nonostante durata infinita.

10.5.2 Sequenze a fase minima e a fase massima

Consideriamo alcune conseguenze importanti delle proprietà P2-P4. Esaminiamo dapprima una sequenza a fase minima con trasformata z della forma

$$X(z) = |A| \frac{\prod_{k=1}^{m_i} (1 - a_k z^{-1})}{\prod_{k=1}^{p_i} (1 - c_k z^{-1})}$$

Chiaramente è

$$x(n) = 0, \quad n < 0 \quad (10.46a)$$

e, dalla proprietà P2,

$$\hat{x}(n) = 0, \quad n < 0 \quad (10.46b)$$

Quindi, per ingressi a fase minima, la sequenza $\hat{x}(n)$ è causale. In questo caso la rappresentazione matematica del sistema D_* può essere notevolmente semplificata. Ricordiamo dal cap. 7 che la trasformata z di una sequenza causale è completamente determinata dalla parte reale della sua trasformata di Fourier. Poiché $x(n)$ e $\hat{x}(n)$ sono entrambe causali, dovremo allora calcolare soltanto

$$\hat{X}_R(e^{j\omega}) = \log |X(e^{j\omega})|$$

per ottenere $\hat{x}(n)$. Ricordando che la trasformata di Fourier inversa di $\hat{X}_R(e^{j\omega})$ è la parte pari di $\hat{x}(n)$, che indichiamo con $c(n)$, si ha

$$c(n) = \frac{\hat{x}(n) + \hat{x}(-n)}{2}$$

Essendo per una sequenza a fase minima $\hat{x}(n) = 0$ per $n < 0$, possiamo scrivere

$$\hat{x}(n) = c(n) \cdot u_+(n) \quad (10.47)$$

dove è

$$u_+(n) = \begin{cases} 0, & n < 0 \\ 1, & n = 0 \\ 2, & n > 0 \end{cases}$$

Queste operazioni sono illustrate in fig. 10.15. La sequenza $c(n)$ è detta il *cepstrum* dell'ingresso $x(n)$ a causa della sua stretta somiglianza con la definizione originale del lavoro di Bogert e altri. È da notare che solo nel caso di ingressi a fase minima (o a fase massima) si può ricavare il cepstrum complesso come in fig. 10.15.

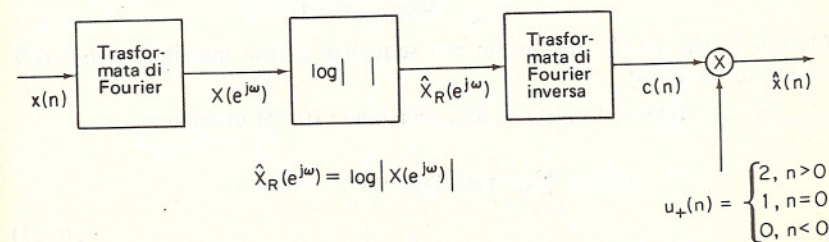


Fig. 10.15 Realizzazione del sistema D_* per un ingresso a fase minima.

Una rappresentazione alternativa si ottiene partendo dall'equazione alle differenze (10.35). Applicando le condizioni espresse nelle (10.46) si ricava

$$\begin{aligned} x(n) &= \sum_{k=0}^n \left(\frac{k}{n}\right) \hat{x}(k)x(n-k), \quad n > 0 \\ &= \hat{x}(n)x(0) + \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{x}(k)x(n-k) \end{aligned} \quad (10.48)$$

Esplicitando rispetto a $\hat{x}(n)$ si giunge alla formula ricorsiva

$$\hat{x}(n) = \begin{cases} 0, & n < 0 \\ \frac{x(n)}{x(0)} - \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{x}(k) \frac{x(n-k)}{x(0)}, & n > 0 \end{cases} \quad (10.49)$$

Si dimostra facilmente (v. probl. 9 di questo capitolo) che il valore di $\hat{x}(0)$ è

$$\hat{x}(0) = \log [A] = \log [x(0)] \quad (10.50)$$

Allora le relazioni (10.49) e (10.50) costituiscono una rappresentazione del sistema D_* per segnali a fase minima. Dalla (10.49) segue anche che quando gli ingressi sono a fase minima il sistema D_* è causale, cioè l'uscita per $n < n_0$ dipende solo dall'ingresso per $n < n_0$, con n_0 arbitrario (v. probl. 15 alla fine del capitolo). Analogamente, le relazioni (10.48) e (10.50) rappresentano il sistema caratteristico inverso D_*^{-1} .

Nel caso di sequenze a fase massima si può svolgere un ragionamento parallelo al precedente. Una sequenza a fase massima non ha poli né zeri interni al cerchio unitario. Allora è

$$x(n) = \hat{x}(n) = 0, \quad n > 0 \quad (10.51)$$

È ancora vero che occorre solo $\log |X(e^{j\omega})|$ per calcolare $\hat{x}(n)$, in quanto è

$$\hat{x}(n) = u_-(n) \cdot c(n)$$

con

$$u_-(n) = \begin{cases} 2, & n < 0 \\ 1, & n = 0 \\ 0, & n > 0 \end{cases} \quad (10.52)$$

Quindi la fig. 10.15 vale anche per sequenze a fase massima se $u_+(n)$ è sostituito da $u_-(n)$.

Se applichiamo la (10.51) alla relazione (10.35) otteniamo

$$\begin{aligned} x(n) &= \sum_{k=n}^0 \left(\frac{k}{n}\right) \hat{x}(k)x(n-k), \quad n < 0 \\ &= \hat{x}(n)x(0) + \sum_{k=n+1}^0 \left(\frac{k}{n}\right) \hat{x}(k)x(n-k) \end{aligned} \quad (10.53)$$

e, risolvendo rispetto a $\hat{x}(n)$,

$$\hat{x}(n) = \begin{cases} \frac{x(n)}{x(0)} - \sum_{k=n+1}^0 \left(\frac{k}{n}\right) \hat{x}(k) \frac{x(n-k)}{x(0)}, & n < 0 \\ \log [x(0)], & n = 0 \\ 0, & n > 0 \end{cases} \quad (10.54)$$

Queste relazioni rappresentano quindi il sistema caratteristico ed il suo inverso per ingressi a fase massima.

La discussione sulle sequenze a fase minima e massima ha una conseguenza interessante per le sequenze di lunghezza finita. Precisamente, nonostante la proprietà P4, possiamo dimostrare che per una sequenza d'ingresso lunga N occorrono solo N campioni di $\hat{x}(n)$ per determinare $x(n)$. Per rendercene conto, consideriamo la trasformata z

$$X(z) = X_{\min}(z) \cdot X_{\max}(z)$$

dove è

$$\begin{aligned} X_{\min}(z) &= A \prod_{k=1}^{m_i} (1 - a_k z^{-1}) \\ X_{\max}(z) &= \prod_{k=1}^{m_o} (1 - b_k z) \end{aligned}$$

La sequenza corrispondente è

$$x(n) = x_{\min}(n) * x_{\max}(n)$$

dove $x_{\min}(n) = 0$ all'esterno dell'intervallo $0 \leq n \leq m_i$ e $x_{\max}(n) = 0$ all'esterno dell'intervallo $-m_o \leq n \leq 0$. La sequenza $x(n)$ è quindi diversa da zero nell'intervallo $-m_o \leq n \leq m_i$. Usando le formule ricorsive prima derivate possiamo scrivere

$$x_{\min}(n) = \begin{cases} 0 & n < 0 \\ e^{\hat{x}(0)} & n = 0 \\ \hat{x}(n)x(0) + \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{x}(k)x_{\min}(n-k) & n > 0 \end{cases} \quad (10.55)$$

e

$$x_{\max}(n) = \begin{cases} 0 & n > 0 \\ 1 & n = 0 \\ \hat{x}(n) + \sum_{k=n+1}^0 \left(\frac{k}{n}\right) \hat{x}(k)x_{\max}(n-k) & n < 0 \end{cases} \quad (10.56)$$

È chiaro che occorrono $m_i + 1$ valori di $\hat{x}(n)$ per calcolare $x_{\min}(n)$ e m_o valori di $\hat{x}(n)$ per calcolare $x_{\max}(n)$. Perciò sono necessari solo $m_i + m_o + 1$

valori della sequenza di lunghezza infinita $\hat{x}(n)$ per ricostruire completamente la sequenza di lunghezza finita $x(n)$.

10.5.3 Poli e zeri sul circolo unitario

Finora non abbiamo considerato poli e zeri sul circolo unitario e vi sono buone ragioni per averlo fatto, sia dal punto di vista teorico che del calcolo. Ricordiamo che nella rappresentazione matematica del sistema caratteristico abbiamo scelto come percorso di integrazione il circolo unitario. Se $X(z)$ ha un polo o uno zero sul circolo unitario, non possiamo associare a $\log[X(z)]$ una regione di convergenza che comprenda il circolo unitario stesso. Si può dimostrare che un fattore $\log(1 - e^{j\theta}e^{-j\omega})$ è esprimibile in serie di Fourier come

$$\log(1 - e^{j\theta}e^{-j\omega}) = - \sum_{n=1}^{\infty} \frac{e^{jn\theta}}{n} e^{-jn\omega}$$

che converge rispetto a qualche criterio. Però la parte reale del logaritmo complesso è infinita e la parte immaginaria discontinua, il che crea delle difficoltà aggiuntive. Per evitarle, scegliamo un diverso percorso C per calcolare $\hat{x}(n)$ da $\log[X(z)]$. In maniera equivalente, possiamo moltiplicare la sequenza d'ingresso per una sequenza esponenziale ottenendo

$$w(n) = \alpha^n x(n)$$

dove α è reale e positivo. La sequenza risultante ha una trasformata z

$$W(z) = X(\alpha^{-1}z)$$

I poli e gli zeri di $X(z)$ sono quindi traslati radialmente di un fattore α^{-1} . È importante notare che se $x(n) = x_1(n) * x_2(n)$, allora si ha

$$W(z) = X(\alpha^{-1}z) = X_1(\alpha^{-1}z) \cdot X_2(\alpha^{-1}z)$$

in modo che risulta

$$w(n) = \alpha^n x_1(n) * \alpha^n x_2(n)$$

Questo equivale a dire che il pesare con un esponenziale una convoluzione dà luogo alla convoluzione di sequenze pesate esponenzialmente.

Oltre a fornire un mezzo per spostare le singolarità di $\log[X(z)]$ dal circolo unitario, l'uso di una funzione peso esponenziale è anche una tecnica utile per trasformare un segnale a fase mista in uno a fase minima oppure a fase massima.

10.6 ALGORITMI PER LA REALIZZAZIONE DEL SISTEMA CARATTERISTICO D_*

Nel par. 10.4 abbiamo dato alcune rappresentazioni matematiche della trasformazione omomorfa D_* , che abbiamo chiamato sistema caratteristico per la convoluzione, e il cui scopo è di trasformare una combinazione mediante convoluzione in una combinazione mediante somma, in modo

che si possa applicare il filtraggio lineare. In tutte queste rappresentazioni era implicita l'ipotesi di univocità e continuità del logaritmo complesso, e in due di esse giocava un ruolo fondamentale la trasformata di Fourier. Se queste rappresentazioni matematiche devono servire come base per realizzare numericamente il sistema D_* , allora è necessario affrontare i problemi connessi con il calcolo della trasformata di Fourier e del logaritmo complesso.

10.6.1 Realizzazione basata sul logaritmo complesso

Il sistema D_* è rappresentato dalle equazioni

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-jn\omega} \quad (10.57a)$$

$$\hat{X}(e^{j\omega}) = \log[X(e^{j\omega})] \quad (10.57b)$$

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) e^{jn\omega} d\omega \quad (10.57c)$$

Per il modo in cui operano i calcolatori numerici, dobbiamo limitarci a considerare sequenze d'ingresso di lunghezza finita e trasformate di Fourier solo per un numero finito di punti. Ciò implica, ovviamente, che si debba fare uso della trasformata di Fourier discreta. Quindi, al posto delle (10.57), usiamo per il calcolo le espressioni seguenti:

$$X(k) = X(e^{j\omega})|_{\omega=(2\pi/N)k} = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn} \quad (10.58a)$$

$$\hat{X}(k) = \log[X(e^{j\omega})]|_{\omega=(2\pi/N)k} = \log[X(k)] \quad (10.58b)$$

$$\hat{x}_p(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k) e^{j(2\pi/N)kn} \quad (10.58c)$$

In base al teorema del campionamento per la trasformata z ricavato nel cap. 3, è chiaro che $\hat{x}_p(n)$ è legata alla $\hat{x}(n)$ desiderata dalla

$$\hat{x}_p(n) = \sum_{k=-\infty}^{\infty} \hat{x}(n + kN) \quad (10.59)$$

Poiché il cepstrum complesso ha in generale durata infinita, $x_p(n)$ sarà una versione con *aliasing* (temporale) di $\hat{x}(n)$; tuttavia, nella proprietà P1 abbiamo notato che $\hat{x}(n)$ in generale tende a zero più rapidamente di una sequenza esponenziale, cosicché è da attendersi che l'approssimazione diventi sempre migliore al crescere di N . Perciò può essere necessario aggiungere zeri a una sequenza d'ingresso affinché il logaritmo complesso sia campionato abbastanza fittamente da evitare un eccessivo *aliasing* nel calcolo del cepstrum complesso.

Scrivendo le relazioni (10.58) e (10.59) abbiamo assunto che $\hat{X}(k)$ rappresenti una versione campionata del logaritmo complesso continuo.

Dobbiamo quindi esaminare come calcolare i campioni di $\arg[X(e^{j\omega})]$ dalla DFT $X(k)$. Un semplice algoritmo è basato sul calcolo di

$$-\pi < \text{ARG}[X(k)] \leq \pi$$

che si ottiene usando le normali funzioni di libreria per l'arcotangente, disponibili su quasi tutti i calcolatori. Questi campioni del valore principale della fase sono poi « srotolati » per ricavare i campioni della curva di fase continua. Consideriamo un ingresso di lunghezza finita con trasformata di Fourier

$$\begin{aligned} X(e^{j\omega}) &= \sum_{n=0}^M x(n)e^{-j\omega n} \\ &= Ae^{-j\omega m_0} \prod_{k=1}^{m_i} (1 - a_k e^{-j\omega}) \prod_{k=1}^{m_o} (1 - b_k e^{j\omega}) \end{aligned} \quad (10.60)$$

dove $|a_k|$ e $|b_k|$ sono minori di uno ed è $M = m_0 + m_i$. In fig. 10.16(a) è rappresentata una curva di fase continua per una sequenza di questo tipo. I punti marcati indicano i campioni di $X(e^{j\omega})$ ad $\omega = (2\pi/N)k$ richiesti per il calcolo di $\hat{x}_p(n)$ (N è assunto pari). La fig. 10.16(b) mostra il valore

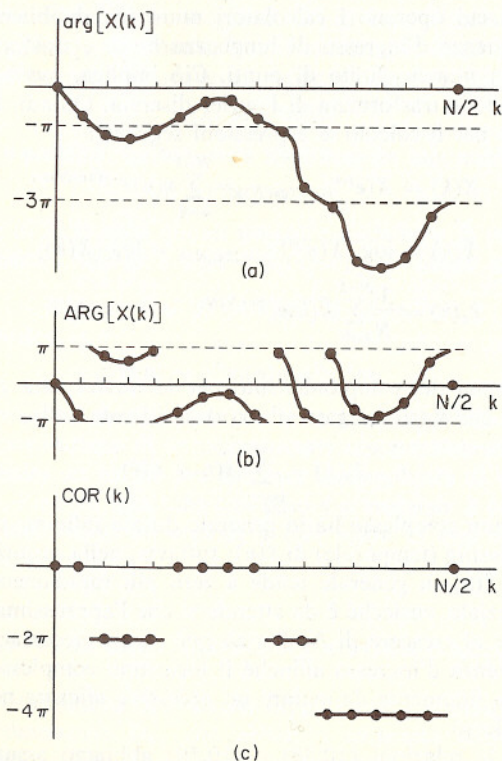


Fig. 10.16 (a) campioni di $\arg[X(e^{j\omega})]$; (b) valore principale di (a); (c) sequenza correttiva per ottenere \arg da ARG .

principale e i suoi campioni ricavati dalla DFT dell'ingresso. Si può vedere che per ottenere i campioni della fase desiderata occorre aggiungere un opportuno multiplo intero di 2π ai campioni del valore principale. Questo multiplo opportuno di 2π può essere ricavato da $\text{ARG}[X(k)]$ se i campioni sono abbastanza ravvicinati da rivelare le discontinuità. Se $\arg[X(e^{j\omega})]$ presenta variazioni rapide ci aspettiamo che $\hat{x}(n)$ tenda a zero meno velocemente che nel caso di variazioni lente. Ma, se $\arg[X(e^{j\omega})]$ varia rapidamente, richiede anche un campionamento più fitto per assicurare la rivelazione delle discontinuità di $\text{ARG}[X(e^{j\omega})]$. Quindi la minimizzazione dell'*aliasing* e la possibilità di ricavare i campioni della curva continua di fase vanno di pari passo. Maggiore è il valore di N , migliore è l'approssimazione dei calcoli. Questa non è in generale una limitazione importante grazie all'esistenza degli algoritmi di FFT: ed effettivamente la scoperta dell'algoritmo di Cooley-Tukey fu determinante nel promuovere le applicazioni dei sistemi omomorfi per la convoluzione.

Un commento finale sul calcolo del logaritmo complesso continuo riguarda il segno di A e la componente a fase lineare dovuta al fattore $e^{-j\omega m_0}$. Il segno di A può essere ricavato facilmente essendo uguale al segno di $X(k)$ per $k = 0$. Il valore di m_0 può essere determinato sommando ad $\text{ARG}[X(k)]$ la sequenza di correzione, poiché si dimostra con facilità dalla (10.60) che è

$$\arg[X(e^{j\pi})] = -m_0\pi$$

Prima di calcolare il cepstrum complesso, viene sottratta dalla fase questa componente a fase lineare, ed il segno di A viene reso positivo.

10.6.2 Realizzazione basata sulla derivata logaritmica

In alternativa al calcolo effettivo del logaritmo complesso, può essere usata la rappresentazione matematica basata sulla derivata logaritmica. In termini della trasformata di Fourier questa rappresentazione è la seguente:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \quad (10.61)$$

$$X'(e^{j\omega}) = -j \sum_{n=-\infty}^{\infty} nx(n)e^{-j\omega n} \quad (10.62)$$

$$\hat{x}(n) = \frac{-1}{2\pi j} \int_{-\pi}^{\pi} \frac{X'(e^{j\omega})}{X(e^{j\omega})} e^{j\omega n} d\omega, \quad n \neq 0 \quad (10.63)$$

$$\hat{x}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(e^{j\omega})| d\omega \quad (10.64)$$

Per sequenze di lunghezza finita ed usando la DFT invece della trasformata di Fourier, queste relazioni diventano:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)kn} = X(e^{j\omega}) \Big|_{\omega=2\pi k/N} \quad (10.65)$$

$$X'(k) = -j \sum_{n=0}^{N-1} nx(n) e^{-j(2\pi/N)kn} = X'(e^{j\omega}) \Big|_{\omega=2\pi k/N} \quad (10.66)$$

$$\hat{x}_{dp}(n) = -\frac{1}{jN} \sum_{k=0}^{N-1} \frac{X'(k)}{X(k)} e^{j(2\pi/N)kn}, \quad 1 \leq n \leq N-1 \quad (10.67)$$

$$\hat{x}_{dp}(0) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| \quad (10.68)$$

dove l'indice d si riferisce all'uso della derivata logaritmica e l'indice p anticipa la periodicità intrinseca nei calcoli della DFT. In questo caso evitiamo il problema di calcolare il logaritmo complesso, al prezzo di un *aliasing* più sensibile, essendo adesso

$$\hat{x}_{dp}(n) = \frac{1}{N} \sum_{k=-\infty}^{\infty} (n + kN) \hat{x}(n + kN) \quad (10.69)$$

Quindi, nell'ipotesi che la curva di fase campionata sia calcolata con sufficiente precisione, ci si aspetta che, per un dato valore di N , $\hat{x}_p(n)$ della (10.58c) sia un'approssimazione a $\hat{x}(n)$ migliore di $\hat{x}_{dp}(n)$ della (10.67).

Per sequenze di lunghezza finita con trasformata di Fourier del tipo (10.60) si può dimostrare che è

$$m_o = \frac{-1}{2\pi j} \int_{-\pi}^{\pi} \frac{X'(e^{j\omega})}{X(e^{j\omega})} d\omega$$

Approssimando questa espressione mediante la DFT inversa si ottiene

$$m_{op} = \frac{-1}{jN} \sum_{k=0}^{N-1} \frac{X'(k)}{X(k)}$$

La quantità m_{op} non sarà, in generale, intera; tuttavia, per valori di N grandi, è lecito aspettarsi che m_{op} tenda a m_o , il numero di zeri di $X(z)$ esterni al cerchio unitario.

10.6.3 Realizzazioni a fase minima

Nel caso particolare di ingressi a fase minima, la rappresentazione matematica si semplifica come in fig. 10.15. Le relazioni su cui si basa la realizzazione sono in questo caso le seguenti:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)kn} \quad (10.70a)$$

$$\hat{X}_R(k) = \log |X(k)| \quad (10.70b)$$

$$c_p(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}_R(k) e^{j(2\pi/N)kn} \quad (10.70c)$$

In questa struttura è il cepstrum a presentare *aliasing*, cioè risulta

$$c_p(n) = \sum_{k=-\infty}^{\infty} c(n + kN)$$

Per calcolare il cepstrum complesso da $c_p(n)$ secondo lo schema di fig. 10.15 scriviamo

$$\hat{x}_{cp}(n) = \begin{cases} c_p(n), & n = 0, N/2 \\ 2c_p(n), & 1 \leq n < N/2 \\ 0, & N/2 < n \leq N-1 \end{cases}$$

Chiaramente è $\hat{x}_{cp}(n) \neq \hat{x}_p(n)$ poiché qui è la parte pari di $\hat{x}(n)$ che subisce *aliasing* e non $\hat{x}(n)$ nel suo complesso. Ciò nonostante, per valori di N grandi, è lecito attendersi che $\hat{x}_{cp}(n)$ differisca molto poco da $\hat{x}(n)$. Analogamente, se $x(n)$ è a fase massima, si otterrebbe un'approssimazione del cepstrum complesso dalla

$$\hat{x}_{cp}(n) = \begin{cases} 0, & 1 \leq n < N/2 \\ c_p(n), & n = 0, N/2 \\ 2c_p(n), & N/2 < n \leq N-1 \end{cases}$$

Nel caso di sequenze a fase minima o a fase massima abbiamo anche a disposizione le formule ricorsive (10.49) - (10.56) come possibili realizzazioni del sistema caratteristico e del suo inverso. Queste relazioni sono molto utili quando la sequenza d'ingresso è molto corta o quando sono richiesti solo pochi valori del cepstrum complesso. Con queste formule non si ha, ovviamente, alcun errore di *aliasing*.

10.7 APPLICAZIONI DELLA DECONVOLUZIONE OMOMORFA

I concetti teorici discussi nei paragrafi 10.4-10.6 sono stati applicati in molti problemi di elaborazione dei segnali. Queste applicazioni possono essere suddivise in due classi: (1) stima di parametri della voce, e (2) eliminazione dei riverberi, che consiste, in senso lato, nella deconvoluzione di due o più segnali di cui uno costituito da una sequenza d'impulsi.

10.7.1 Stima di parametri della voce

Cominciamo con il presentare un modello relativo alla generazione del segnale voce. Poi faremo vedere come si possono utilizzare i risultati teorici dei paragrafi precedenti per stimare i parametri di questo modello.

La voce viene prodotta dall'eccitazione di un condotto acustico, chiamato *tratto vocale*, che termina da una parte con le labbra e dall'altra con la glottide. Fintantoché rimane in una configurazione fissa, il tratto vocale può essere modellato come un sistema lineare tempo-invariante la cui uscita è la convoluzione della risposta all'impulso del tratto vocale con la

forma d'onda di eccitazione. La voce è generata in tre modi fondamentali. I *suoni vocalici* sono prodotti eccitando il tratto vocale con un flusso d'aria fatto di impulsi quasi periodici e causato dalla vibrazione delle corde vocali. I *suoni fricativi* sono prodotti formando un restringimento nel tratto vocale e facendovi passare di forza l'aria, in modo che si genera una turbolenza e quindi un'eccitazione assimilabile a rumore. I *suoni plosivi* vengono prodotti chiudendo completamente il tratto vocale, esercitandovi una pressione e riaprendolo di colpo. Tutti e tre questi modi di generare la voce (sorgenti) corrispondono a un'eccitazione a larga banda del tratto vocale, che può essere modellato come un filtro che varia lentamente nel tempo e modifica con la sua risposta in frequenza lo spettro dell'eccitazione. Il tratto vocale è caratterizzato dalle sue frequenze naturali (dette *formanti*), che corrispondono a risonanze nelle sue caratteristiche di trasmissione.

Se ammettiamo che le sorgenti di eccitazione e la forma del tratto vocale siano approssimativamente indipendenti, un accettabile modello è quello di fig. 10.17. In questo modello a tempo discreto si assume che i campioni della voce siano l'uscita di un filtro numerico tempo-variante che

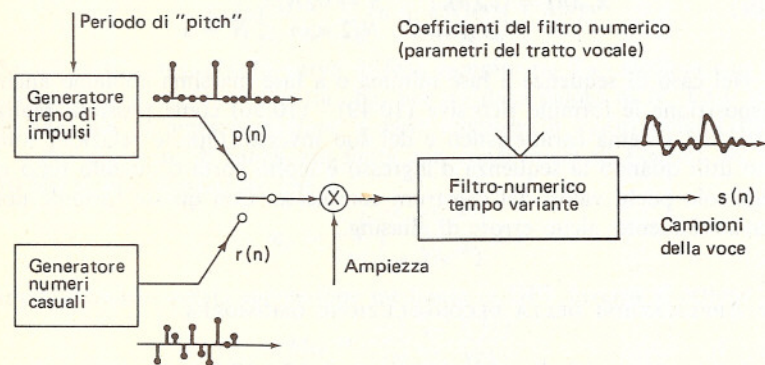


Fig. 10.17 Modello di generazione della voce.

approssima le proprietà trasmissive del tratto vocale. Poiché durante l'emissione della voce il tratto vocale cambia forma piuttosto lentamente, è ragionevole assumere che questo filtro abbia caratteristiche costanti in un intervallo di tempo dell'ordine di 10 ms. In ognuno di questi intervalli di tempo il filtro numerico può quindi essere caratterizzato da una risposta all'impulso o da una risposta in frequenza o da un insieme di coefficienti relativi ad un filtro con risposta all'impulso di durata infinita. Precisamente, per i suoni vocalici (esclusi quelli nasali), la funzione di trasferimento del filtro numerico consiste in una componente relativa al tratto vocale

$$V(z) = \frac{A}{\prod_{k=1}^p (1 - c_k z^{-1})(1 - c_k^* z^{-1})}, \quad |c_k| < 1 \quad (10.71)$$

dove i c_k corrispondono alle frequenze naturali del tratto vocale, e in una componente addizionale

$$G(z) = B \prod_{k=1}^{m_1} (1 - a_k z^{-1}) \prod_{k=1}^{m_0} (1 - b_k z^{-1}) \quad (10.72)$$

che tiene conto del fatto che gli impulsi di durata finita provenienti dalla glottide non sono impulsi ideali. Quindi la funzione di trasferimento del filtro numerico di fig. 10.17 è

$$H_v(z) = G(z)V(z)$$

Questo filtro viene eccitato da un treno di impulsi $p(n)$, in cui la distanza tra gli impulsi corrisponde al periodo fondamentale (detto « pitch ») della voce.

Per i suoni non vocalici la teoria della propagazione delle onde acustiche suggerisce che la caratteristica di trasmissione del tratto vocale presenta sia zeri che poli. Si può assumere allora il modello

$$H_u(z) = \frac{A \prod_{k=1}^m (1 - \alpha_k z^{-1})(1 - \alpha_k^* z^{-1})}{\prod_{k=1}^p (1 - c_k z^{-1})(1 - c_k^* z^{-1})} \quad (10.73)$$

dove è $|c_k| < 1$. In questo caso il sistema è pilotato da una sequenza di rumore casuale $r(n)$. In entrambi i casi, vocalico o non vocalico, un controllo di ampiezza regola l'intensità dell'ingresso al filtro numerico.

Se si assume che in un intervallo di tempo breve il sistema è quindi il modello di generazione della voce rimanga costante, allora è possibile stimarne i parametri applicando la deconvoluzione omomorfa [11]. Infatti, ogni breve segmento di voce costituito di suoni vocalici può essere pensato come la convoluzione

$$s(n) = p(n) * g(n) * v(n), \quad 0 \leq n \leq L - 1$$

Per minimizzare inoltre gli effetti delle discontinuità all'inizio e alla fine dell'intervallo, si usa una « finestra dati » $w(n)$ che moltiplica $s(n)$ in modo che l'ingresso al sistema omomorfo è

$$x(n) = s(n)w(n)$$

Se $w(n)$ varia lentamente rispetto al termine $g(n) * v(n)$, possiamo scrivere

$$x(n) \simeq p_w(n) * [g(n) * v(n)] \quad (10.74a)$$

dove è

$$p_w(n) = w(n) \cdot p(n) \quad (10.74b)$$

Esaminiamo i contributi di ogni componente della (10.74a) al cepstrum complesso. Si può ragionevolmente assumere che, nel breve intervallo di tempo della finestra, $p(n)$ sia un treno di impulsi equispaziati

$$p(n) = \sum_{k=0}^{M-1} \delta(n - kn_0)$$

in modo che è

$$p_w(n) = \sum_{k=0}^{M-1} w(kn_0) \delta(n - kn_0)$$

dove abbiamo fatto l'ipotesi che la finestra abbracci M impulsi. Se definiamo una sequenza.

$$w_{n_0}(k) = \begin{cases} w(kn_0) & k = 0, 1, \dots, M-1 \\ 0 & \text{altrove} \end{cases}$$

allora la trasformata di Fourier di $p_w(n)$ è

$$\begin{aligned} P_w(e^{j\omega}) &= \sum_{k=0}^{M-1} w(kn_0) e^{-j\omega kn_0} \\ &= W_{n_0}(e^{j\omega n_0}) \end{aligned} \quad (10.75)$$

Perciò $P_w(e^{j\omega})$ ed anche $\hat{P}_w(e^{j\omega})$ sono periodiche con periodo $2\pi/n_0$. Il cepstrum complesso di $p_w(n)$ è

$$\hat{p}_w(n) = \hat{w}_{n_0}\left(\frac{n}{n_0}\right) \quad n = 0, \pm n_0, \pm 2n_0, \dots$$

Quindi la periodicità del logaritmo complesso dà luogo nel cepstrum complesso ad impulsi spaziali di n_0 campioni. Se la sequenza $w_{n_0}(n)$ è a fase minima, allora $\hat{p}_w(n)$ sarà nulla per $n < 0$. In caso contrario, $\hat{p}_w(n)$ presenterà degli impulsi ad intervalli di n_0 campioni per n sia positivi che negativi. In ogni caso si ha contributo di $\hat{p}_w(n)$ a $x(n)$ nella regione $|n| \geq n_0$.

Il cepstrum complesso di $v(n)$ può essere ricavato dal logaritmo complesso di $V(z)$:

$$\hat{V}(z) = \log [A] - \sum_{k=1}^p \{ \log [1 - c_k z^{-1}] + \log [1 - c_k^* z^{-1}] \} \quad (10.76)$$

Si vede facilmente da questa espressione che è

$$\hat{v}(n) = \begin{cases} 0 & n < 0 \\ \log [A] & n = 0 \\ \frac{1}{n} \sum_{k=1}^p [(c_k)^n + (c_k^*)^n] & n > 0 \end{cases} \quad (10.77)$$

o, se scriviamo $c_k = |c_k| e^{j\phi_k}$,

$$\hat{v}(n) = \sum_{k=1}^p \frac{|c_k|^n}{n} 2 \cos \phi_k n \quad n > 0 \quad (10.78)$$

L'impulso prodotto nella glottide, $g(n)$, è di durata finita e si assume che non sia, in generale, a fase minima. Allora $g(n)$ può essere rappresentata come la convoluzione di una sequenza a fase minima con una sequenza a fase massima:

$$g(n) = g_{\min}(n) * g_{\max}(n) \quad (10.79)$$

Il contributo al cepstrum complesso $\hat{x}(n)$ dovuto a $g(n)$ è

$$\hat{g}(n) = \begin{cases} \hat{g}_{\min}(n) & 0 \leq n \\ \hat{g}_{\max}(n) & n < 0 \end{cases} \quad (10.80)$$

dove ci si può aspettare, in base alla discussione precedente, che il maggior contributo di $\hat{g}(n)$ a $\hat{x}(n)$ sia nella regione intorno a $n = 0$.

In generale le componenti del cepstrum complesso, $\hat{v}(n)$ e $\hat{g}(n)$, tendono a zero piuttosto rapidamente, in modo che per valori abbastanza grandi di n_0 i contributi del tratto vocale e dell'impulso della glottide non si sovrappongono a $\hat{p}_w(n)$ ². In altri termini, nel logaritmo complesso le componenti dovute al tratto vocale e alla glottide variano lentamente e quelle dovute al « pitch » variano rapidamente. Ciò è illustrato in fig. 10.18. La fig. 10.18(a) mostra un segmento di voce pesato con una finestra di Hamming, e la fig. 10.18(b) presenta il logaritmo complesso della trasformata di Fourier discreta di fig. 10.18(a)³. Si notino la componente rapidamente variabile — quasi periodica — dovuta a $p_w(n)$, e le componenti a variazione lenta dovute a $v(n)$ e $g(n)$. Queste proprietà si manifestano nel cepstrum complesso di fig. 10.18(c) attraverso gli impulsi a multipli di circa 8 ms (il periodo della voce in ingresso), dovuti a $\hat{p}_w(n)$, e attraverso i campioni nell'intervallo $|nT| < 5$ ms, che attribuiamo a $\hat{v}(n)$ e a $\hat{g}(n)$.

Se vogliamo separare le componenti del segnale voce, il suggerimento che si ricava dalla discussione precedente è di filtrare il logaritmo complesso con un passa-basso per ottenere $v(n) * g(n)$ e con un passa-alto per ricavare $p_w(n)$. Un esempio è presentato in fig. 10.19. La fig. 10.19(a) mostra un segmento di una vocale. Dopo averlo pesato con una finestra di Hamming, il cepstrum complesso risulta quello di fig. 10.19(b). Se il cepstrum complesso viene moltiplicato per la sequenza

$$l(n) = \begin{cases} 0 & |n| \leq 40 \\ 1 & |n| > 40 \end{cases}$$

e il risultato filtrato dal sistema caratteristico inverso D_*^{-1} , l'uscita che si ottiene è quella di fig. 10.19(c)⁴. D'altro lato, per riottenere l'impulso $v(n) * g(n)$ moltiplichiamo il cepstrum complesso per la sequenza

$$l(n) = \begin{cases} 1 & |n| \leq 40 \\ 0 & |n| > 40 \end{cases}$$

L'uscita di D_*^{-1} in questo caso è mostrata in fig. 10.19(d). La fig. 10.19(e) presenta il risultato della convoluzione della forma d'onda di fig. 10.19(d)

² Per voce campionata a 10 kHz, il campo di variazione tipico per il periodo di « pitch » è $40 < n_0 < 150$.

³ In tutte le figure di questo paragrafo i campioni di tutte le sequenze sono collegati con segmenti di retta per comodità di rappresentazione.

⁴ La frequenza di campionamento è di 10 kHz, in modo che 40 campioni corrispondono a 4 ms.

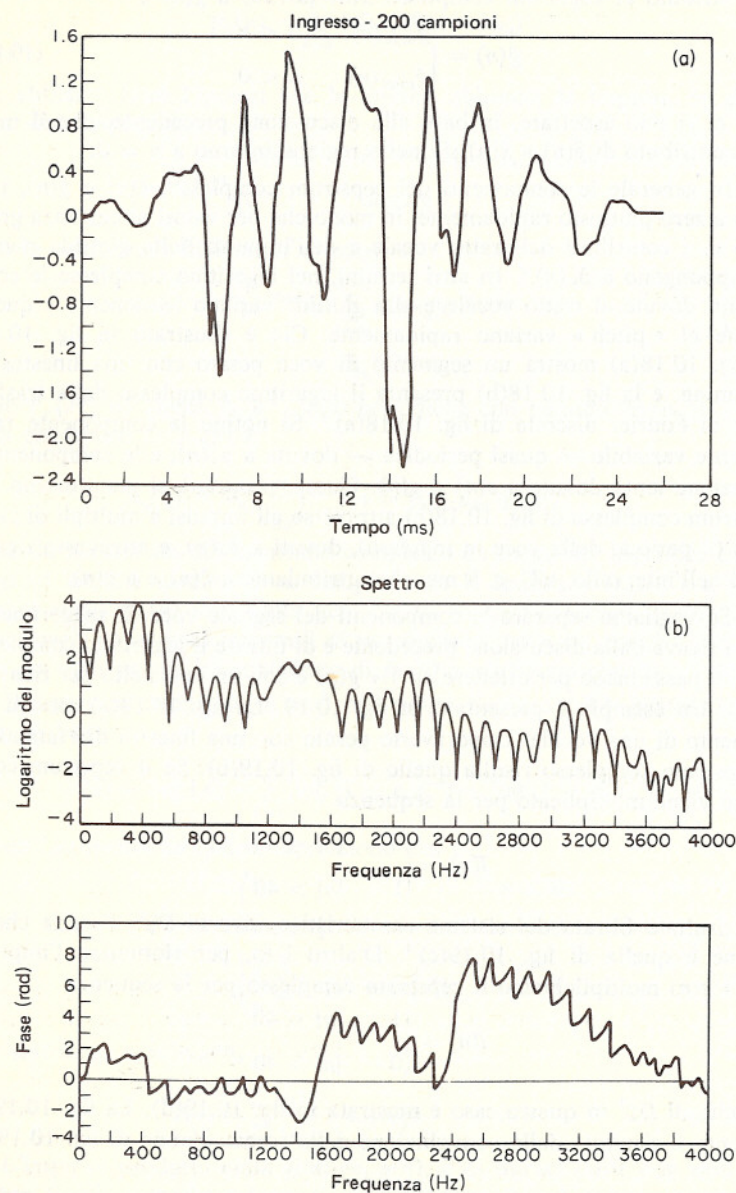


Fig. 10.18 (a) segmento di voce pesato con una finestra di Hamming; (b) logaritmo complesso della trasformata di (a); (c) cepstrum complesso di (a).

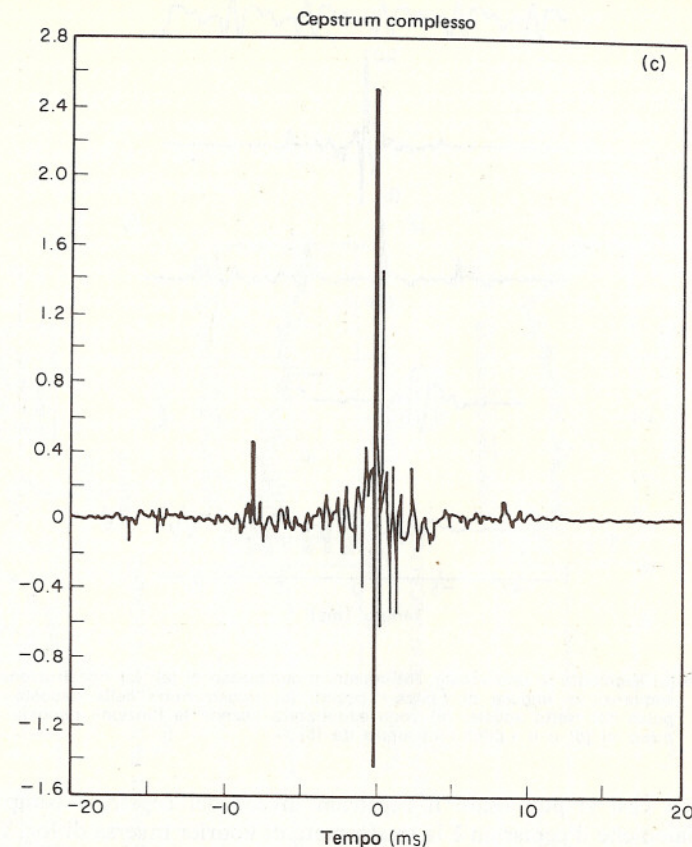


Fig. 10.18(c)

con una sequenza d'impulsi fatta di impulsi unitari di uguale ampiezza localizzati in istanti corrispondenti a quelli dei picchi di fig. 10.19(c).

La discussione precedente ha mostrato che la deconvoluzione omomorfa può essere applicata con successo per *separare* le componenti del segnale voce. Tuttavia, in molte applicazioni di analisi della voce ci interessa soltanto *stimare* alcuni parametri della voce piuttosto che ricostruire le effettive forme d'onda componenti. Per esempio, può essere sufficiente decidere se un particolare segmento di voce è vocalico o non vocalico e poi, se vocalico, stimare il periodo di « pitch » o l'involuppo dello spettro

$$\log |V(e^{j\omega})G(e^{j\omega})|$$

e, se non vocalico, stimare lo spettro

$$\log |H_u(e^{j\omega})|$$

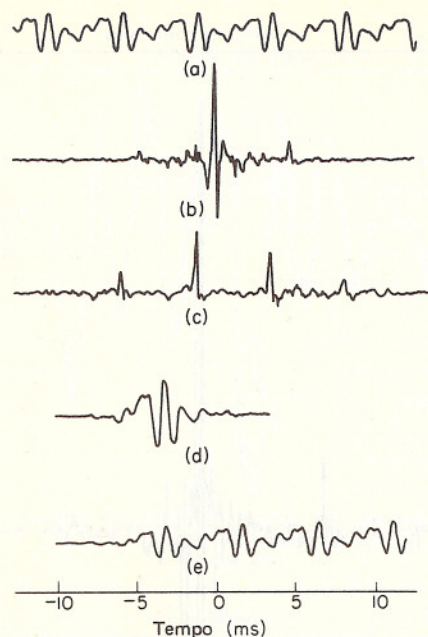


Fig. 10.19 (a) segmento di una vocale; (b) cepstrum complesso di (a); (c) ricostruzione della sequenza di impulsi di «pitch» pesati; (d) ricostruzione della risposta all'impulso del tratto vocale; (e) voce sintetizzata usando la funzione risposta all'impulso di (d) e il «pitch» misurato da (b).

In questi casi si può usare il cepstrum invece del cepstrum complesso. Ricordiamo che il cepstrum è la trasformata di Fourier inversa di $\log|X(e^{j\omega})|$ e quindi risulta

$$c(n) = \frac{1}{2}[\hat{x}(n) + \hat{x}(-n)]$$

Perciò si può dire che la parte di $c(n)$ vicina all'origine dell'asse tempo corrisponde alle componenti lentamente variabili di $\log|X(e^{j\omega})|$ che sono determinate dalla configurazione del tratto vocale; inoltre, nel caso di suoni vocalici, la componente pari di $p_w(n)$ dovrebbe contenere impulsi negli stessi istanti di $\hat{p}_w(n)$. Tutto ciò è rappresentato in fig. 10.20. La fig. 10.20(a) illustra i calcoli relativi alla stima dei parametri della voce. La fig. 10.20(b) mostra un risultato tipico per suoni vocalici. Il segmento di voce da analizzare, già pesato con una finestra, è indicato con A, $\log|X(k)|$ è indicato con C e il cepstrum $c(n)$ con D. La posizione del picco nel cepstrum a circa 8 ms fornisce una misura del periodo di «pitch» proprio di questo segmento di voce. L'involuppo dello spettro, ottenuto moltiplicando $c(n)$ per una finestra che ne mantiene solo la parte intorno all'origine ($|n| < 40$) e calcolandone poi la DFT, è indicato con E ed è sovrapposto a $\log|X(k)|$. Per i suoni non vocalici la situazione, illustrata in fig. 10.20(c), è del tutto

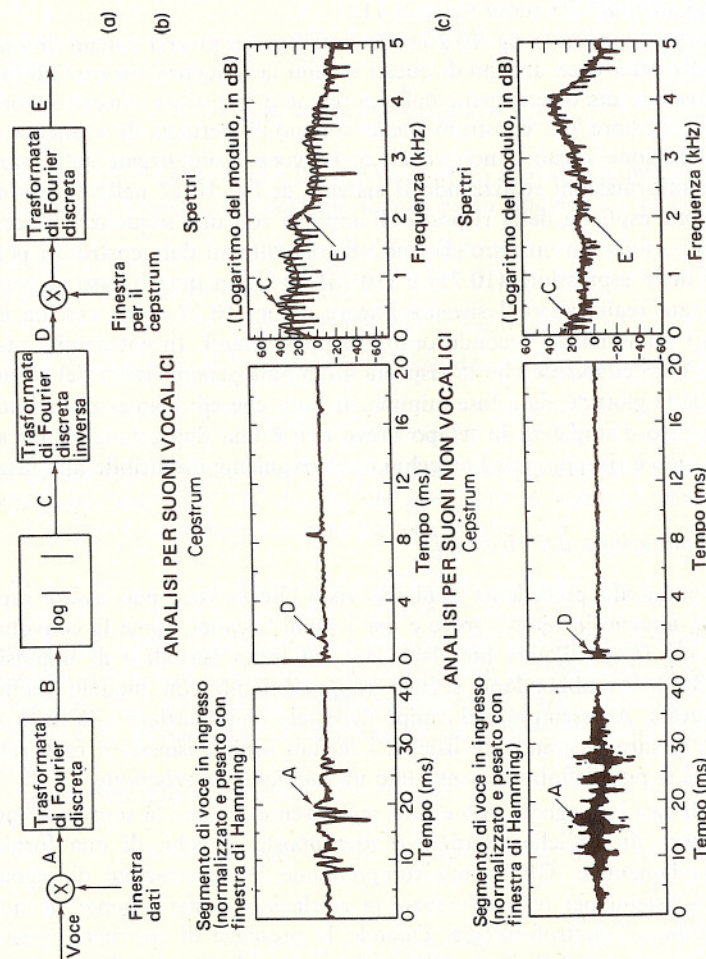


Fig. 10.20 (a) sistema per l'analisi omomorfa della voce; (b) analisi per suoni vocalici; (c) analisi per suoni non vocalici.

analoga, con la differenza che la natura di rumore dell'eccitazione produce ora in $\log|X(k)|$ una componente casuale rapidamente variabile. Quindi le componenti del cepstrum vicino all'origine corrispondono alla funzione di trasferimento del tratto vocale; però, essendo le variazioni rapide di $\log|X(k)|$ non periodiche, non esiste un grosso picco come nel caso dei suoni vocalici. Allora il cepstrum costituisce un metodo eccellente per trovare se un segmento di voce è vocalico o non vocalico e per stimare il periodo fondamentale dei suoni vocalici [12].

I metodi illustrati in fig. 10.20 sono stati usati in diversi sistemi di analisi e sintesi della voce. In uno di questi sistemi la sequenza risposta all'impulso viene ricavata direttamente dalla porzione di cepstrum intorno all'origine [13]. Sempre dal cepstrum viene stimato il periodo di « pitch » e presa la decisione vocalico/non vocalico. La voce è sintetizzata sulla base di queste informazioni realizzando il sistema di fig. 10.17 nella forma di convoluzione esplicita della risposta all'impulso con una sequenza di eccitazione opportuna. In un altro sistema vengono stimati dal cepstrum i poli e gli zeri delle espressioni (10.71) e (10.73) [14]. In questo caso la voce è sintetizzata realizzando il sistema lineare di fig. 10.17 come cascata di risuonatori numerici del secondo ordine tempo-varianti. In entrambi i casi si assume implicitamente che la risposta all'impulso complessiva del tratto vocale e della glottide sia a fase minima. Il fatto che ciò mantenga validità al solo spettro d'ampiezza in tempo breve non è una limitazione significativa in quanto è risaputo che l'orecchio è relativamente insensibile alla fase.

10.7.2 Eliminazione dei riverberi

Nel paragrafo precedente abbiamo visto che la voce può essere rappresentata, almeno in tempo breve e per i suoni vocalici, come la convoluzione di una forma d'onda impulsiva con un treno periodico di impulsi. Nel mondo fisico abbondano segnali rappresentabili con modelli molto simili a questo. Ad esempio, nel campo delle telecomunicazioni o là dove si effettuano misure di grandezze fisiche, i segnali sono trasmessi o registrati in quel che si può definire in senso lato un *ambiente riverberante*.

In tali casi un segnale può essere rappresentato come la somma di un certo numero di repliche ritardate e sovrapposte, o echi, di una forma d'onda fondamentale. Gli esempi comprendono le registrazioni di segnali fonici, i sistemi per teleconferenze, la rivelazione radar e sonar, le misure sismiche e l'elettrofisiologia. Quando la presenza di riverberi è vista come una distorsione della forma d'onda fondamentale, si cerca di riottenere quest'ultima. In altri casi, lo scopo dell'elaborazione è invece quello di determinare proprio la configurazione degli echi, in quanto caratterizzante una struttura o un processo fisico. Nel seguito di questo paragrafo discuteremo alcuni esempi di applicazione della deconvoluzione omomorfa a segnali del tipo sopra descritto.

Cominciamo considerando una sequenza che sia la somma di repliche ritardate e scalate in ampiezza di una sequenza $s(n)$, cioè

$$x(n) = s(n) + \sum_{k=1}^M \alpha_k s(n - n_k) \quad (10.81)$$

dove è $0 < n_1 < n_2 < \dots < n_M$. Questo segnale può essere rappresentato come la convoluzione

$$x(n) = s(n) * p(n) \quad (10.82a)$$

dove è

$$p(n) = \delta(n) + \sum_{k=1}^M \alpha_k \delta(n - n_k) \quad (10.82b)$$

Come semplice esempio che illustri l'uso di un sistema omomorfo per questa classe di segnali, consideriamo il caso di un singolo eco, cioè

$$p(n) = \delta(n) + \alpha_1 \delta(n - n_1) \quad (10.83)$$

La trasformata di Fourier di $x(n)$ è

$$X(e^{j\omega}) = S(e^{j\omega})(1 + \alpha_1 e^{-j\omega n_1}) \quad (10.84)$$

Allora il contributo al logaritmo complesso dovuto alla sequenza di impulsi è

$$\hat{P}(e^{j\omega}) = \log(1 + \alpha_1 e^{-j\omega n_1}) \quad (10.85)$$

In questo caso semplice $\hat{P}(e^{j\omega})$ è periodico con periodo $2\pi/n_1$ e perciò ci aspettiamo che $\hat{p}(n)$ sia diverso da zero solo a multipli interi di n_1 . Se è $|\alpha_1| < 1$, si dimostra facilmente che risulta

$$\hat{p}(n) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\alpha_1^k}{k} \delta(n - kn_1) \quad (10.86)$$

Quindi, se $\hat{S}(e^{j\omega})$ varia lentamente rispetto alle variazioni di $\hat{P}(e^{j\omega})$, si può pensare di separare queste due componenti con un filtro lineare frequenza-invariante. Ad esempio, se vogliamo riottenere $p(n)$, possiamo usare un filtro che fa passare solo le componenti del cepstrum complesso a tempi lunghi.

Nel caso generale risulta

$$p(n) = \delta(n) + \sum_{k=1}^M \alpha_k \delta(n - n_k) \quad (10.87)$$

e

$$P(e^{j\omega}) = 1 + \sum_{k=1}^M \alpha_k e^{-j\omega n_k} \quad (10.88)$$

Se gli echi sono equispaziati, cioè $n_k = kn_1$, allora si è visto nel par. 10.7.1 che il cepstrum complesso ha la stessa forma che nel caso di un'eco singola. In generale, tuttavia, non ci possiamo aspettare che gli echi siano equispaziati e quindi il cepstrum complesso di $p(n)$ consisterà di impulsi ad istanti che sono funzioni complicate dei ritardi originari. Tuttavia, nel

caso particolare di $p(n)$ a fase minima, sappiamo che è $\hat{p}(n) = 0$ per $n < 0$. Inoltre si può dimostrare [8] che è $\hat{p}(n) = 0$ per $n < n_1$, dove n_1 è il ritardo minimo, e il cepstrum complesso per $n > n_1$ sarà fatto di impulsi che si verificano agli istanti

$$n_l = \sum_{k=1}^M l n_k, \quad l = 0, 1, 2, \dots \quad (10.89)$$

con ampiezze decrescenti al crescere di n . Alla luce della discussione del par. 10.5.3 è chiaro che un treno di impulsi a fase non minima può essere reso a fase minima moltiplicandolo per una funzione peso esponenziale. Vale a dire che, per β abbastanza piccolo, la sequenza $\beta^n p(n)$ è certamente a fase minima. In molti casi conviene moltiplicare per una funzione peso esponenziale perché così si ottiene una certa separazione tra le componenti del cepstrum complesso dovute a $\beta^n s(n)$ e a $\beta^n p(n)$.

Avendo presenti queste proprietà del cepstrum complesso di un treno d'impulsi, consideriamo alcuni esempi dell'uso del filtraggio omomorfo per separare le componenti di una convoluzione della forma (10.82).

Echi nei segnali vocali. In molti canali di comunicazione il segnale vocale o voce può essere distorto a causa di echi o riverberi. Essendo il segnale vocale per sua natura continuo, la forma d'onda deve essere elaborata scomponendola in tratti di dimensione opportuna ed i segmenti che risultano in uscita devono essere ricomposti per formare la sequenza d'uscita complessiva. La fig. 10.18(c) mostra il cepstrum complesso di un segmento di voce. Se nella (10.82) $s(n)$ è un segnale vocale, il cepstrum complesso di un segmento di $x(n)$ conterrà degli impulsi dovuti a $p(n)$ purché la durata della finestra sia maggiore di n_M , che è il ritardo massimo. Se il ritardo più breve, n_1 , è maggiore del massimo periodo di « pitch » (circa 15 ms), il contributo di $p(n)$ non si sovrapporrà in modo significativo al cepstrum complesso del segnale vocale. Un esempio è presentato in fig. 10.21(a).

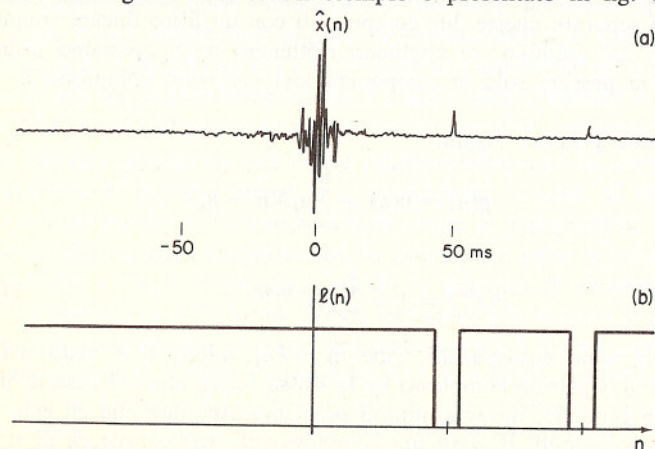


Fig. 10.21 (a) Cepstrum complesso di un segmento di segnale vocale con un'eco a 50 ms; (b) caratteristica del filtro usato per sopprimere l'eco.

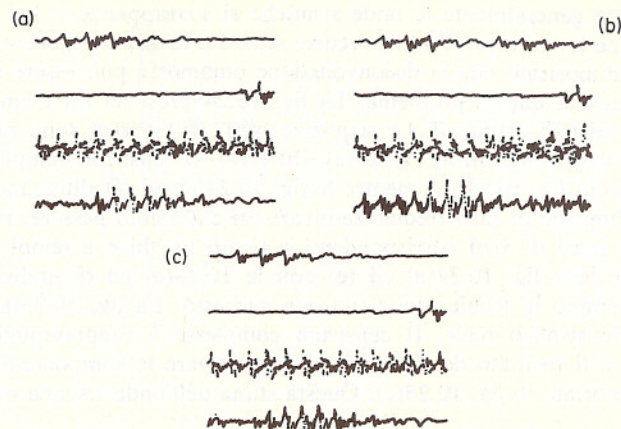


Fig. 10.22 Esempio di soppressione d'eco con filtraggio omomorfo: (a) 410 ms di voce campionata a 10 kHz, con le quattro tracce che rappresentano, dall'alto in basso, segmenti adiacenti lunghi 102.5 ms; (b) stessa voce di (a) con un'eco a 50 ms; (c) voce di (b) elaborata per eliminare l'eco.

In questo caso la voce, campionata a 10 kHz, è stata ritardata e sommata a sé stessa, generando così un segnale

$$x(n) = s(n) + \alpha_1 s(n - n_1) \quad (10.90)$$

Per riottenere $s(n)$ da $x(n)$ dobbiamo eliminare i contributi al cepstrum complesso dovuti all'eco. Si può ottenere questo usando un filtro invariante in frequenza con la caratteristica mostrata in fig. 10.21(b). Il successo di questo tipo di elaborazione è illustrato in fig. 10.22: la fig. 10.22(a) mostra $s(n)$, la 10.22(b) mostra $x(n)$ e nella fig. 10.22(c) è presentata l'uscita del sistema D_{*}^{-1} per il filtro «a pettine» di fig. 10.21(b). In questo caso la voce è stata elaborata a tratti lunghi 2048 campioni e le singole uscite ottenute sono state ricombinate per formare l'uscita complessiva. I dettagli dell'operazione di ricombinazione sono descritti in [8].

Segnali sismici. Le espressioni (10.82) costituiscono anche un utile modello per i segnali sismici. In questo caso un'esplosione genera un impulso di energia sismica che si propaga attraverso la terra e subisce riflessioni in corrispondenza delle superfici di separazione tra i vari strati della crosta terrestre. La fig. 10.23 presenta un modello per i segnali sismici, dove $p(n)$ è una sequenza d'impulsi che contiene l'informazione sulla struttura della crosta terrestre e l'onda sismica $s(n)$ dipende dalla natura dell'eccitazione e dalla dispersione incontrata nella propagazione.

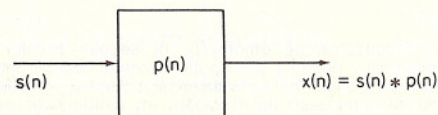


Fig. 10.23 Semplice modello per segnali sismici.

Poiché generalmente le onde sismiche si sovrappongono nel tempo e mascherano la struttura di $p(n)$, occorre separare le due componenti. Ulrych [15] ha dimostrato che la deconvoluzione omomorfa può essere applicata con successo a questo problema. La fig. 10.24 presenta un esempio ottenuto con segnali artificiali. Le sequenze $p(n)$, $s(n)$ e $x(n)$ sono rappresentate, rispettivamente, in fig. 10.24(a), (b) e (c). Il cepstrum complesso è mostrato in fig. 10.24(d), mentre le fig. 10.24(e) ed (f) illustrano il risultato dell'impiego di filtri frequenza-invarianti che fanno passare, rispettivamente, le parti di $\hat{x}(n)$ corrispondenti a tempi lunghi e a tempi brevi. Il confronto delle fig. 10.24(a) ed (e) con le 10.24(c) ed (f) indica che in questo esempio la tecnica funziona con successo. La fig. 10.25(a) mostra un segnale sismico reale. Il cepstrum complesso è rappresentato in fig. 10.25(b), e il risultato del filtraggio che fa passare le componenti a tempi brevi è riportato in fig. 10.25(c). Questa stima dell'onda sismica può essere

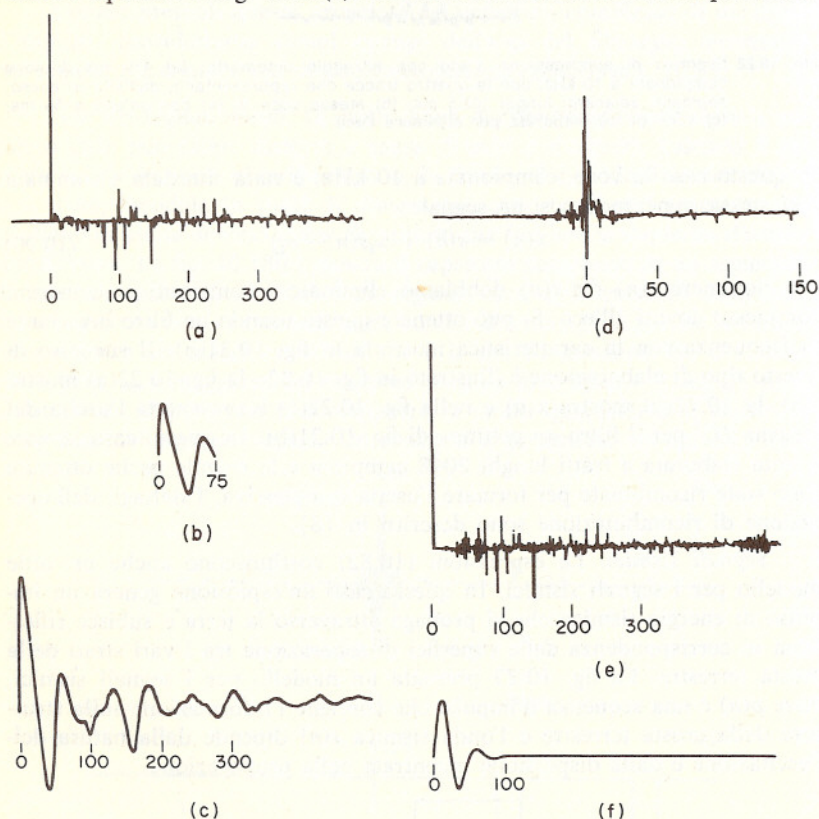


Fig. 10.24 Esempio di deconvoluzione omomorfa di segnali sismici ottenuto con forme d'onda artificiali: (a) risposta all'impulso teorica della crosta presso Leduc, nell'Alberta (da O. Jensen); (b) onda sismica ipotetica; (c) sismogramma sintetizzato; (d) cepstrum complesso del tracciato (c) moltiplicato per una funzione peso esponenziale con $\alpha = 0.985$; (e) uscita del passa-alto; (f) uscita del passa-basso (da Ulrych [15]).

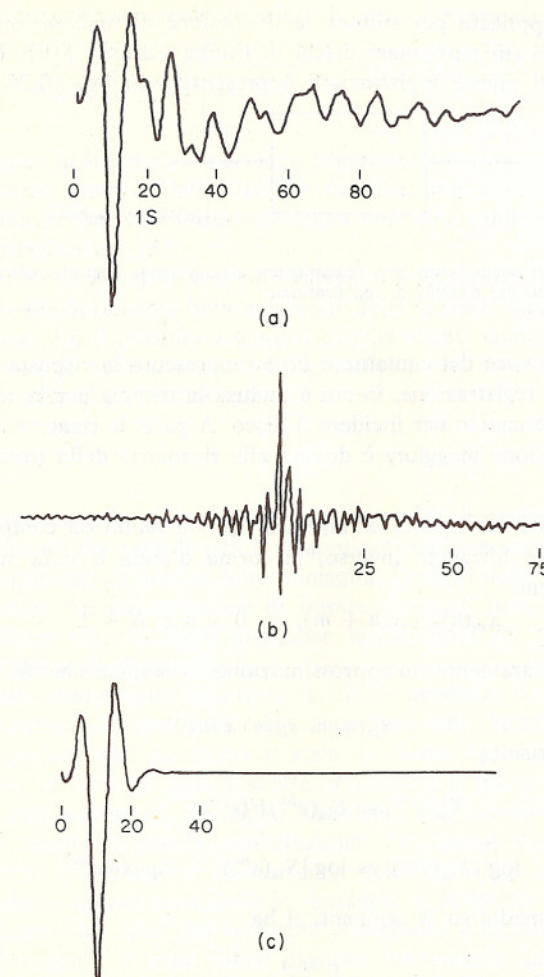


Fig. 10.25 Esempio di deconvoluzione omomorfa di un evento sismico reale registrato a distanza: (a) registrazione, fatta nel 1968 a Leduc nell'Alberta, di un evento sismico verificatosi nel Venezuela; (b) cepstrum complesso di (a) dopo la moltiplicazione per una funzione peso esponenziale con $\alpha = 0.985$; (c) stima dell'onda sismica ottenuta con un filtraggio passa-basso di (b) (da Ulrych [15]).

utile per avere informazioni sulle proprietà di attenuazione e di dispersione del cammino di trasmissione. In entrambi gli esempi è stata usata una funzione peso esponenziale.

10.7.3 Ripristino di registrazioni discografiche

Una ben nota tecnica di riduzione del rumore consiste essenzialmente nel fare la media di un gran numero di segnali in cui la forma d'onda desiderata sia sempre la stessa e invece il rumore sia diverso. Questa tec-

nica è stata applicata per stimare la distorsione di registrazioni fatte con metodi acustici (in particolare dischi di Enrico Caruso) [16]. Un modello semplificato di queste registrazioni è presentato in fig. 10.26, dove $s(n)$

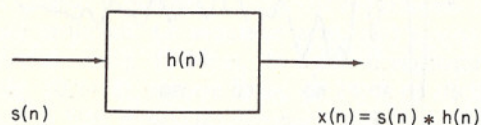


Fig. 10.26 Modello semplificato per registrazioni discografiche distorte dalla risposta all'impulso del sistema di registrazione.

rappresenta la voce del cantante e $h(n)$ rappresenta la risposta all'impulso del sistema di registrazione, in cui è inclusa la tromba per la registrazione e il sistema meccanico per incidere il disco. A parte le rigature della superficie, la distorsione maggiore è dovuta alle risonanze della tromba di registrazione.

Per stimare la risposta all'impulso $h(n)$ in modo da compensarne gli effetti mediante filtraggio inverso, la forma d'onda è stata suddivisa in parti come segue

$$x_m(n) = x(n + m), \quad 0 \leq n \leq N - 1$$

Anche se è chiaramente un'approssimazione, si assume che sia

$$x_m(n) \approx s_m(n) * h(n)$$

in modo che risulta

$$X_m(e^{j\omega}) \approx S_m(e^{j\omega})H(e^{j\omega})$$

e

$$\log |X_m(e^{j\omega})| \approx \log |S_m(e^{j\omega})| + \log |H(e^{j\omega})|$$

Prendendo la media su M segmenti si ha

$$\frac{1}{M} \sum_{m=0}^{M-1} \log |X_m(e^{j\omega})| = \frac{1}{M} \sum_{m=0}^{M-1} \log |S_m(e^{j\omega})| + \log |H(e^{j\omega})|$$

Il termine

$$\frac{1}{M} \sum_{m=0}^{M-1} \log |S_m(e^{j\omega})| \approx \log |S(e^{j\omega})|$$

è una stima del logaritmo dello spettro di potenza a lungo termine della voce o del canto, e se ne può ricavare una buona approssimazione da registrazioni moderne che presentano una distorsione minima. Una stima $H_e(e^{j\omega})$ della risposta in frequenza della tromba si ottiene allora da

$$\log |H_e(e^{j\omega})| = \frac{1}{M} \sum_{m=0}^{M-1} \log |X_m(e^{j\omega})| - \overline{\log |S(e^{j\omega})|}$$

Un filtro inverso che compensi gli effetti di $h(n)$ si ricava da

$$|H_e^{-1}(e^{j\omega})| = \begin{cases} \frac{1}{|H_e(e^{j\omega})|}, & |\omega| \leq \omega_p \\ 0, & \omega_s \leq |\omega| \leq \pi \end{cases} \quad (10.91)$$

dove la risposta in frequenza si attenua linearmente tra ω_p e ω_s . La trasformata di Fourier inversa della (10.91) fornisce una risposta all'impulso con fase nulla $h_e^{-1}(n)$ che viene convoluta con $x(n)$ usando un algoritmo di convoluzione veloce [16].

Nonostante la quantità di approssimazioni fatte, la tecnica descritta ha prodotto dei miglioramenti impressionanti della qualità soggettiva dei dischi di Caruso. Ci si possono attendere miglioramenti analoghi nell'elaborazione di voce o musica registrata in ambienti con riverberi come, per esempio, nel caso delle teleconferenze.

SOMMARIO

In questo capitolo abbiamo discusso una classe di tecniche di elaborazione non lineare dei segnali e la loro applicazione in numerosi campi, tra cui il miglioramento di qualità delle immagini, l'analisi della voce e l'esplorazione sismica. La presentazione di questa classe di tecniche ci ha fornito l'occasione per illustrare numerose applicazioni di risultati teorici ricavati nel corso di questo libro.

Abbiamo considerato dapprima la classe generale dei sistemi omomorfi, concentrandoci poi sulle due sottoclassi che hanno trovato più numerose applicazioni. La prima è stata la classe dei sistemi omomorfi moltiplicativi, di cui sono state descritte in dettaglio le proprietà e l'applicazione alla elaborazione di immagini. Abbiamo poi considerato la classe dei sistemi omomorfi per la convoluzione. Per questi ultimi, numerosi concetti teorici importanti si basano sulla definizione del logaritmo complesso. Abbiamo quindi esaminato in dettaglio le proprietà dell'uscita del sistema caratteristico (cioè del cepstrum complesso) e alcuni algoritmi per la sua realizzazione. È stata infine discussa brevemente l'applicazione di queste idee all'elaborazione della voce, alla soppressione di echi nei segnali vocali, all'analisi di segnali sismici ed al ripristino di registrazioni discografiche.

BIBLIOGRAFIA

1. A. V. Oppenheim, "Superposition in a Class of Nonlinear Systems," *Tech. Rept. 432*, Research Laboratory of Electronics, MIT, Cambridge, Mass., Mar. 1965.
2. A. V. Oppenheim, "Generalized Superposition," *Inform. Control*, Vol. 11, Nos. 5-6, Nov.-Dec. 1967, pp. 528-536.
3. A. V. Oppenheim, R. W. Schaffer, and T. G. Stockham, Jr., "Nonlinear Filtering of Multiplied and Convolved Signals," *Proc. IEEE*, Vol. 56, No. 8, Aug. 1968, pp. 1264-1291.

4. A. V. Oppenheim, "Generalized Linear Filtering," Chapter 8 in *Digital Processing of Signals*, B. Gold and C. M. Rader, McGraw-Hill Book Company, New York, 1969.
5. T. G. Stockham, Jr., "The Application of Generalized Linearity to Automatic Gain Control," *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, June 1968, pp. 267-270.
6. T. G. Stockham, Jr., "Image Processing in the Context of a Visual Model," *Proc. IEEE*, Vol. 60, No. 7, July 1972, pp. 828-842.
7. A. V. Oppenheim, "Nonlinear Filtering of Convolved Signals," *Quart. Progr. Rept. 80*, Research Laboratory of Electronics, MIT, Cambridge, Mass., Jan. 1966, pp. 168-175.
8. R. W. Schafer, "Echo Removal by Discrete Generalized Linear Filtering," *Tech. Rept. 466*, MIT Research Laboratory of Electronics, MIT, Cambridge, Mass., Feb. 1969. Also Ph.D. Thesis, Department of Elec. Engineering, MIT, Feb. 1968.
9. B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Sappe Cracking," *Proc. Symp. Time Series Analysis*, M. Rosenblatt, Ed., New York, John Wiley & Sons, Inc., New York, 1963, pp. 209-243.
10. J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed., Springer-Verlag, New York, 1972.
11. A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, No. 2, June 1968, pp. 221-226.
12. A. M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Amer.*, Vol. 41, Feb. 1967, pp. 293-309.
13. A. V. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," *J. Acoust. Soc. Amer.*, Vol. 45, Feb. 1969, pp. 458-465.
14. R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," *J. Acoust. Soc. Amer.*, Vol. 47, No. 2, Pt. 2, Feb. 1970, pp. 634-648.
15. T. J. Ulrych, "Application of Homomorphic Deconvolution to Seismology," *Geophys.*, Vol. 36, No. 4, Aug. 1971, pp. 650-660.
16. T. G. Stockham, Jr., "Restoration of Old Acoustic Recordings by Means of Digital Signal Processing," Preprint, 41st Convention, Audio Engineering Society, New York, Oct. 1971.

PROBLEMI

1. Ognuno dei sistemi seguenti opera una trasformazione omomorfa. L'operazione d'ingresso è quella indicata. Trovare l'operazione d'uscita.

Trasformazione $T[x(n)]$	Operazione d'ingresso
$y(n) = T[x(n)] = 2x(n)$	Addizione
$y(n) = T[x(n)] = 2x(n)$	Moltiplicazione
$X(z) = T[x(n)] = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$	Addizione
$X(z) = T[x(n)] = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$	Convoluzione
$X(z) = T[x(n)] = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$	Moltiplicazione
$y(n) = T[x(n)] = x^2(n)$	Moltiplicazione

$y(n) = T[x(n)] = x(n) $	Moltiplicazione
$y(n) = e^{x(n)}$	Addizione
$y(n) = e^{x(n)}$	Moltiplicazione

2. Due sistemi omomorfi, H_1 e H_2 , sono posti in cascata. H_1 è omomorfo con operazioni di ingresso e di uscita, rispettivamente, la moltiplicazione e la convoluzione. H_2 è omomorfo con operazioni di ingresso e di uscita, rispettivamente, la convoluzione e l'addizione. Dimostrare che il sistema complessivo è omomorfo, con la moltiplicazione come operazione d'ingresso e l'addizione come operazione d'uscita.
3. Si consideri la classe di sistemi omomorfi con la moltiplicazione come operazione d'ingresso e di uscita. Dimostrare che se l'ingresso $x(n)$ vale uno per tutti gli n , allora anche l'uscita $y(n)$ vale uno per tutti gli n .
4. Trovare quale dei seguenti sistemi non può essere omomorfo, con la moltiplicazione come operazione d'ingresso e d'uscita:
 - (a) $y(n) = 3x(n)$.
 - (b) $y(n) = x^2(n)$.
 - (c) $y(n) = [1/x(n)][x(n) - x(n-1)]$.
 - (d) $y(n) = |x(n)|$.
 - (e) $y(n) = x(n)/x(n-1)$.
5. Consideriamo la classe dei sistemi omomorfi con la convoluzione come operazione d'ingresso e di uscita. Dimostrare che se l'ingresso è $x(n) = \delta(n)$, allora anche l'uscita è $y(n) = \delta(n)$.
6. $\hat{x}(n)$ è il cepstrum complesso di $x(n)$. Nell'elenco seguente, la prima colonna riguarda proprietà relative a $\hat{x}(n)$, mentre la seconda riguarda proprietà relative a $x(n)$. Trovare, per ogni proprietà della prima colonna, la proprietà corrispondente nella seconda. Assumere sempre che $x(n)$ sia reale. Ogni proprietà della seconda colonna può essere usata una sola volta.

(1) $\hat{x}(n)$ reale	(a) $x(n) = -x(-n)$
(2) $\hat{x}(n) = -\hat{x}(-n)$	(b) $x(n) = x(-n)$
(3) $\hat{x}(n) = 0, n < 0$	(c) $x(n)$ reale
	(d) $x(n) = 0, n < 0$
	(e) $\sum_{n=-\infty}^{\infty} x^2(n) = 1$
	(f) $\sum_{n=-\infty}^{\infty} x(n) = 1/\sqrt{2\pi}$
7. $x_1(n)$ e $x_2(n)$ rappresentano due sequenze e $\hat{x}_1(n)$ e $\hat{x}_2(n)$ i loro cepstrum complessi. Se è $x_1(n) * x_2(n) = \delta(n)$, trovare la relazione tra $\hat{x}_1(n)$ e $\hat{x}_2(n)$.
8. Il cepstrum complesso $\hat{x}(n)$ di una sequenza $x(n)$ è stato definito in modo che se è

$$\hat{X}(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} \hat{x}(n)e^{-j\omega n}$$

e

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x(n)e^{-j\omega n} = |X(e^{j\omega})| e^{j\theta(\omega)}$$

allora risulta

$$\hat{X}(e^{j\omega}) = \log X(e^{j\omega}) = \log |X(e^{j\omega})| + j\theta(\omega)$$

dove $\theta(\omega)$ è una funzione continua, dispari e periodica di ω .

Si definisce sequenza a fase minima quella il cui cepstrum complesso è nullo per $n < 0$, e sequenza a fase massima quella il cui cepstrum complesso è nullo per $n > 0$.

Consideriamo due sequenze $x_1(n)$ e $x_2(n)$ con trasformate $X_1(e^{j\omega})$ e $X_2(e^{j\omega})$ rispettivamente. $x_1(n)$ è a fase minima e $x_2(n)$ a fase massima. Se è $|X_1(e^{j\omega})| = |X_2(e^{j\omega})|$, determinare la relazione che esiste tra $x_1(n)$ e $x_2(n)$.

9. Sia $x(n)$ una sequenza a fase minima e $\hat{x}(n)$ il suo cepstrum complesso. Usare il teorema del valore iniziale (probl. 16 del cap. 2) per dimostrare che è $\hat{x}(0) = \log[x(0)]$. Varrebbe lo stesso risultato se $x(n)$ non fosse a fase minima?
10. Sia $x(n)$ una sequenza a fase massima e $\hat{x}(n)$ il suo cepstrum complesso. Dimostrare che è $\hat{x}(0) = \log[x(0)]$.
11. Consideriamo una sequenza $x(n)$ con cepstrum complesso $\hat{x}(n)$ e con trasformata $X(z)$ espressa nella forma

$$X(z) = \frac{A \prod_{k=1}^{m_i} (1 - a_k z^{-1}) \prod_{k=1}^{m_o} (1 - b_k z)}{\prod_{k=1}^{p_i} (1 - c_k z^{-1}) \prod_{k=1}^{p_o} (1 - d_k z)}$$

dove $|a_k|, |b_k|, |c_k|$ e $|d_k|$ sono tutti minori di uno ed A è un numero reale e positivo. Dimostrare che è $\hat{x}(0) = \log A$.

12. Sia $\hat{x}(n)$ il cepstrum complesso di $x(n)$. Definiamo una sequenza $e(n)$ come

$$e(n) = \begin{cases} x(n/N), & n = KN, K = 0, \pm 1, \pm 2, \dots \\ 0, & \text{altrove} \end{cases}$$

Dimostrare che il cepstrum complesso di $e(n)$, che chiamiamo $\hat{e}(n)$, è dato da

$$\hat{e}(n) = \begin{cases} \hat{x}(n/N), & n = KN, K = 0, \pm 1, \pm 2, \dots \\ 0, & \text{altrove} \end{cases}$$

13. La formula (10.49) rappresenta una relazione ricorsiva tra $x(n)$ e $\hat{x}(n)$ quando $x(n)$ è a fase minima. Usare la (10.49) per generare ricorsivamente il cepstrum complesso della sequenza $x(n) = a^n u(n)$.
14. La formula (10.54) rappresenta una relazione ricorsiva tra $x(n)$ e $\hat{x}(n)$ quando $x(n)$ è a fase massima. Usare la (10.54) per generare ricorsivamente il cepstrum complesso della sequenza $x(n)$ data da

$$\begin{aligned} x(0) &= 1 \\ x(-1) &= -a \\ x(n) &= 0, \quad n \neq 0, -1 \end{aligned}$$

15. La formula (10.49) rappresenta una relazione ricorsiva tra una sequenza $x(n)$ ed il suo cepstrum complesso $\hat{x}(n)$. Usando la (10.49) dimostrare che il sistema caratteristico D_* è un sistema causale per ingressi a fase minima, cioè che per ingressi a fase minima $\hat{x}(n)$ per $n < n_0$ dipende solo da $x(n)$ per $n < n_0$, dove n_0 è arbitrario.

11. STIMA DELLO SPETTRO DI POTENZA

11.0 INTRODUZIONE

Uno dei settori applicativi di rilievo per le tecniche di elaborazione numerica dei segnali fin qui considerate, e in particolare per la trasformata di Fourier veloce, è quello della stima delle funzioni di autocovarianza e di densità spettrale di potenza di una sequenza casuale. La necessità di stimare lo spettro di potenza si presenta in numerose situazioni, come la misura dello spettro di rumore per il progetto di filtri lineari ottimi, la rivelazione di segnali a banda stretta mascherati da rumore a larga banda e la stima dei parametri di un sistema lineare usando il rumore come eccitazione.

I fondamenti matematici delle tecniche di stima dello spettro di potenza rientrano nell'argomento più generale della teoria della stima. In pratica, tuttavia, accade generalmente che le tecniche di stima ottima, come quella basata sulla massima verosimiglianza, richiedono più informazione sul segnale di quanta ne sia di solito disponibile. Per questo motivo le stime dello spettro di potenza solitamente usate hanno una notevole base empirica, e inoltre ogni tecnica ha pregi e difetti, di modo che è impossibile definire in generale il metodo migliore.

Lo scopo di questo capitolo è di fornire un'introduzione breve ed elementare alla stima dello spettro di potenza. Si vuole soprattutto dare un'idea della metodologia con cui viene effettuata la stima dello spettro e del ruolo che vi possono giocare alcune tecniche di elaborazione numerica dei segnali che abbiamo discusso in precedenza. Non sarà tuttavia immediato passare dall'introduzione a livello elementare che faremo alla comprensione approfondita della miriade di compromessi, alternative e tecniche con cui si ha a che fare nella pratica della stima di spettri.

La bibliografia fondamentale sulla stima dello spettro comprende i libri di Bartlett [1], Blackman e Tukey [2], Grenander e Rosenblatt [3] e Hannan [4]. Tra i testi più recenti citiamo quelli di Jenkins e Watts [5] e di Koopmanns [6]. Nella trattazione che segue attingeremo molto da questi lavori di base. Cominceremo con una breve introduzione ad alcuni concetti della teoria generale della stima applicati al caso della stima di medie di un processo aleatorio. Tratteremo poi l'applicazione di questi concetti di base al problema della stima delle sequenze di autocorrelazione o di autocovarianza di un processo aleatorio stazionario. Esamineremo quindi

diversi metodi di stima dello spettro di potenza, dedicando particolare attenzione ad alcune difficoltà che spesso si incontrano nell'applicazione di tecniche standard. Infine, discuteremo l'applicazione di alcuni metodi di elaborazione numerica dei segnali, presentati in capitoli precedenti, alla stima di sequenze di correlazione e spettri di potenza.

11.1 PRINCIPI FONDAMENTALI DI TEORIA DELLA STIMA

Nel cap. 8 abbiamo discusso il concetto di processo aleatorio e la sua caratterizzazione per mezzo di medie delle variabili che lo costituiscono. Quando si caratterizza empiricamente un segnale modellandolo come un processo aleatorio, occorre spesso stimare delle medie del processo che costituisce il modello a partire da una singola sequenza campione del processo stesso, cioè una sequenza $x(n)$ che si assume essere una realizzazione di un processo casuale definito dall'insieme delle variabili aleatorie $\{x_n\}$. Inoltre, per rendere possibile il calcolo delle stime, occorre ricavarle da un segmento finito della sequenza campione $x(n)$. Il fatto di ricavare le stime che interessano da un segmento finito, $x(n)$ per $0 \leq n \leq N-1$, di una singola sequenza campione $x(n)$ è giustificato se si considerano processi ergodici, cioè processi aleatori per i quali le medie d'insieme sono uguali alle medie temporali. Consideriamo ad esempio un processo casuale per cui sia

$$m_x = E[x_n] = \int_{-\infty}^{\infty} x p_x(x) dx, \quad \text{qualsiasi } n \quad (11.1)$$

Supponiamo inoltre che sia

$$m_x = \langle x_n \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x_n \quad (11.2)$$

e, per ogni sequenza campione del processo aleatorio,

$$m_x = \langle x(n) \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n) \quad (11.3)$$

Allora è corretto affermare che la quantità

$$\hat{m}_x = \frac{1}{N} \sum_{n=0}^{N-1} x(n)$$

rappresenta una stima sufficientemente accurata di m_x purché N sia « abbastanza grande ». La parte della statistica che si occupa di questo tipo di problemi è chiamata *teoria della stima* e ne vogliamo illustrare brevemente i concetti fondamentali, prima di esaminare le tecniche pratiche di calcolo delle medie di un processo aleatorio. In un contesto più generale di quello considerato qui, esistono molti problemi che si possono porre a proposito di una sequenza aleatoria. Per esempio, se la sequenza casuale è stata prodotta eccitando un sistema lineare discreto con rumore bianco, può interessare stimare i parametri del sistema lineare. Un altro possibile scopo del-

l'analisi può essere semplicemente quello di decidere se il processo è bianco o no. Si può poi voler caratterizzare il processo stimando parametri come la media, la varianza, la sequenza di autocovarianza, o la densità spettrale di potenza, ed è proprio alla stima di questi parametri che dedicheremo interamente la nostra attenzione nei prossimi paragrafi.

Consideriamo un processo aleatorio stazionario $\{x_n\}$, $-\infty < n < \infty$. Il suo valore medio, m_x , è definito dalla (11.1) e la sua media temporale è definita dalla (11.2). Assumiamo anche che la media temporale di ogni sequenza campione, espressa dalla (11.3), sia uguale a m_x . La varianza del processo aleatorio è definita come

$$\sigma_x^2 = E[(x_n - m_x)^2] = \langle (x_n - m_x)^2 \rangle \quad (11.4)$$

La sequenza di autocovarianza è definita come

$$\gamma_{xx}(m) = E[(x_n - m_x)(x_{n+m}^* - m_x^*)] = \langle (x_n - m_x)(x_{n+m}^* - m_x^*) \rangle \quad (11.5)$$

e la densità spettrale di potenza come

$$P_{xx}(\omega) = \sum_{m=-\infty}^{\infty} \gamma_{xx}(m) e^{-j\omega m} \quad (11.6)$$

Per stimare un parametro « α » del processo aleatorio si ha in generale a disposizione un segmento finito di una singola sequenza campione, cioè N valori $x(n)$, $0 \leq n \leq N-1$. La stima $\hat{\alpha}$ del parametro α è dunque funzione delle variabili aleatorie x_n , $0 \leq n \leq N-1$, cioè

$$\hat{\alpha} = F[x_0, x_1, \dots, x_{N-1}]$$

e quindi anche $\hat{\alpha}$ è una variabile aleatoria. La funzione densità di probabilità di $\hat{\alpha}$ verrà indicata con $p_{\hat{\alpha}}(\hat{\alpha})$. La forma funzionale e l'andamento di $p_{\hat{\alpha}}(\hat{\alpha})$ dipenderanno dalla scelta dello stimatore $F[\]$ e dalle densità di probabilità delle variabili aleatorie x_n . È ragionevole definire uno stimatore « buono » se è elevata la probabilità che la stima sia prossima ad α . In base a questo criterio è evidente che lo stimatore 2 di fig. 11.1 è migliore dello stimatore 1, in quanto la densità di probabilità dello stimatore 2 è più concentrata attorno al valore vero α .

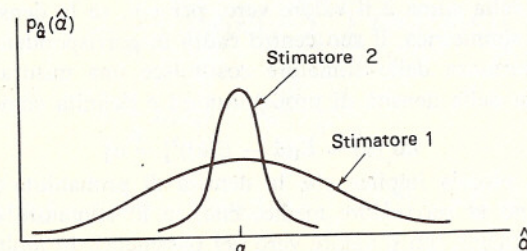


Fig. 11.1 Funzioni densità di probabilità di due stimatori.

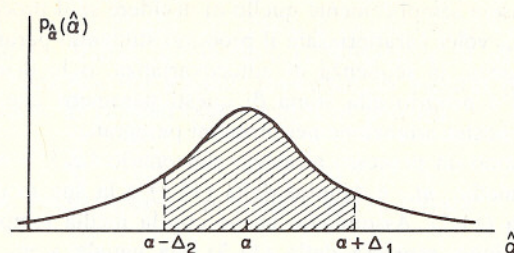


Fig. 11.2 Limiti di confidenza per uno stimatore.

Un modo di caratterizzare la concentrazione della funzione densità di probabilità di uno stimatore è attraverso il concetto di intervallo di confidenza. Ad esempio, per la funzione densità di probabilità rappresentata in fig. 11.2, l'area sotto la curva nell'intervallo $\alpha - \Delta_2 \leq \hat{\alpha} \leq \alpha + \Delta_1$ rappresenta la probabilità che la stima sia compresa tra questi due limiti. Quindi, indicando con $(1 - \beta)$ quell'area, risulta

$$\text{Probabilità } [-\Delta_2 \leq (\hat{\alpha} - \alpha) \leq \Delta_1] = (1 - \beta).$$

Dunque, se per un particolare stimatore abbiamo trovato, ad esempio, che per $\Delta_1 = \Delta_2 = 0.1$ l'area $(1 - \beta)$ è uguale a 0.95, allora possiamo dire che la stima sarà entro l'intervallo $(-0.1, +0.1)$ intorno al valore vero al livello di confidenza del 95 %.

In generale ha senso dire che per uno stimatore buono la funzione densità di probabilità $p_a(\hat{\alpha})$ deve essere stretta e concentrata attorno al valore vero, e si possono usare proprio questi criteri per confrontare tra loro stimatori diversi. Le proprietà degli stimatori che sono usate comunemente come base per il confronto sono infatti la polarizzazione e la varianza. La polarizzazione di uno stimatore è definita come il valore vero del parametro meno il valore atteso della stima, cioè

$$\text{polarizzazione} = \alpha - E[\hat{\alpha}] \triangleq B \quad (11.7)$$

Uno stimatore è non polarizzato se B è nulla: ciò significa allora che il valore atteso della stima è il valore vero, per cui, se la densità di probabilità $p_a(\hat{\alpha})$ è simmetrica, il suo centro cadrà in corrispondenza del valore vero α . La varianza dello stimatore costituisce una misura significativa della larghezza della densità di probabilità ed è definita come

$$\text{var}[\hat{\alpha}] = E[(\hat{\alpha} - E[\hat{\alpha}])^2] = \sigma_{\hat{\alpha}}^2 \quad (11.8)$$

Una varianza piccola implica che la densità di probabilità $p_a(\hat{\alpha})$ è concentrata attorno al suo valore medio, che, se lo stimatore è anche non polarizzato, coincide con il valore vero del parametro. In molti casi il confronto tra due stimatori è complicato dal fatto che quello con la polarizzazione minore ha la varianza maggiore o viceversa. Di conseguenza, è a

volte opportuno considerare l'errore quadratico medio associato ad uno stimatore, definito come

$$\text{errore quadratico medio} = E[(\hat{\alpha} - \alpha)^2] = \sigma_{\hat{\alpha}}^2 + B^2 \quad (11.9)$$

Uno stimatore si dice *consistente* se la polarizzazione e la varianza tendono entrambe a zero al crescere del numero di osservazioni.

Per esemplificare i concetti fin qui esposti, consideriamo un processo aleatorio con funzione densità di probabilità gaussiana, cioè

$$p_{x_n}(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-m_x)^2/2\sigma_x^2}$$

Assumeremo anche che le variabili casuali $\{x_n\}$ siano statisticamente indipendenti, in modo che, in particolare, x_0, x_1, \dots, x_{N-1} sono reali e statisticamente indipendenti. Una classe di stimatori molto usata è quella delle *stime a massima verosimiglianza*. La stima a massima verosimiglianza è basata sulla probabilità congiunta relativa agli N valori delle osservazioni come funzione del parametro da stimare. La stima a massima verosimiglianza è quel valore del parametro per cui è massima la probabilità di ottenere proprio i valori osservati. Per quanto riguarda il problema affrontato qui, è ben noto [7] che la stima a massima verosimiglianza del valor medio m_x di un processo aleatorio gaussiano è la media campione, definita come

$$\text{media campione} = \hat{m}_x = \frac{1}{N} \sum_{i=0}^{N-1} x_i \quad (11.10)$$

Questa è allora una scelta per lo stimatore del parametro m_x . Essendo \hat{m}_x una somma pesata di variabili casuali gaussiane indipendenti, anche la densità di probabilità $p_{\hat{m}_x}(\hat{m}_x)$ è gaussiana [7]. Se $p_{\hat{m}_x}(\hat{m}_x)$ è gaussiana, allora è caratterizzata completamente dalla polarizzazione e dalla varianza dello stimatore. Il valore atteso di \hat{m}_x è uguale al valore atteso di x_n e di conseguenza la polarizzazione è nulla. Per ottenere la varianza della media campione occorre calcolare

$$\begin{aligned} E[\hat{m}_x^2] &= \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} E[x_i x_j] \\ &= \frac{1}{N^2} \left[\sum_{i=0}^{N-1} E[x_i^2] + \sum_{i=0}^{N-1} \sum_{j=0, j \neq i}^{N-1} E[x_i] \cdot E[x_j] \right] \\ &= \frac{1}{N} E[x_n^2] + m_x^2 \frac{N-1}{N} \end{aligned}$$

Perciò risulta

$$\text{var}[\hat{m}_x] = E[\hat{m}_x^2] - \{E[\hat{m}_x]\}^2 = \frac{1}{N} (E[x_n^2] - m_x^2) = \frac{1}{N} \sigma_x^2 \quad (11.11)$$

La formula (11.11) ci dice allora che, al crescere del numero di osservazioni, la varianza della media campione diminuisce: essendo nulla la polarizzazione, si conclude che la media campione è uno stimatore consistente.

Se è noto il valor medio ma occorre stimare la varianza, allora lo stimatore a massima verosimiglianza è

$$\hat{\sigma}_x^2 = \frac{1}{N} \sum_{i=0}^{N-1} (x_i - m_x)^2 \quad (11.12)$$

È immediato verificare che questo stimatore della varianza è consistente. Esso richiede però che sia conosciuto il parametro m_x . Se invece devono essere stimate sia la media che la varianza, allora la stima a massima verosimiglianza della media è la media campione, come prima, e la stima a massima verosimiglianza della varianza è la varianza campione definita come

$$\hat{\sigma}_x^2 = \frac{1}{N} \sum_{i=0}^{N-1} (x_i - \hat{m}_x)^2 \quad (11.13)$$

dove \hat{m}_x è la media campione. La formula (11.13) differisce dalla (11.12) perché in questa compare il valore vero m_x , mentre nell'altra si usa la sua stima. Per trovare la polarizzazione della varianza campione espressa dalla (11.13), possiamo innanzitutto calcolare il valore atteso di $\hat{\sigma}_x^2$. È dunque

$$\begin{aligned} E[\hat{\sigma}_x^2] &= \frac{1}{N} \sum_{i=0}^{N-1} (E[x_i^2] + E[\hat{m}_x^2] - 2E[x_i \hat{m}_x]) \\ &= \frac{1}{N} \sum_{i=0}^{N-1} E[x_i^2] + \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} E[x_i x_j] - \frac{2}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} E[x_i x_j] \\ &= \frac{N-1}{N} E[x_i^2] - \frac{N-1}{N} m_x^2 \\ &= \frac{N-1}{N} \sigma_x^2 \end{aligned} \quad (11.14)$$

Di conseguenza, il valor medio della varianza campione non coincide con la varianza e quindi la varianza campione è polarizzata. Però, quando N diventa molto grande, il valore atteso della varianza campione tende alla varianza, per cui questa stima è asintoticamente non polarizzata. Per calcolare la varianza della varianza campione assumiamo, per comodità, che il processo sia a media nulla, in modo che risulta

$$v = \hat{\sigma}_x^2 = \frac{1}{N} \sum_{i=0}^{N-1} x_i^2$$

Allora si ha

$$\begin{aligned} E[v^2] &= \frac{1}{N^2} \sum_{i=1}^N \sum_{r=1}^N E[x_i^2 x_r^2] \\ &= \frac{1}{N^2} [NE[x_n^4] + N(N-1)\{E[x_n^2]\}^2] \\ &= \frac{1}{N} [E[x_n^4] + (N-1)\{E[x_n^2]\}^2] \end{aligned}$$

Si vede facilmente che è

$$E[v] = E[x_n^2]$$

da cui segue

$$\begin{aligned} \text{var} [\hat{\sigma}_x^2] &= E[v^2] - (E[v])^2 \\ &= \frac{1}{N} \{E[x_n^4] - (E[x_n^2])^2\} \end{aligned} \quad (11.15)$$

Dalle espressioni (11.14) e (11.15) notiamo dunque che la varianza campione è una stima consistente.

La discussione precedente aveva lo scopo di illustrare il tipo di analisi usata per descrivere le proprietà degli stimatori. L'obiettivo di questa analisi è di dare un'idea della precisione della stima e di come la precisione dipenda dal numero di campioni che intervengono nella stima.

Per calcolare i limiti di confidenza per gli stimatori dobbiamo conoscere la distribuzione probabilistica delle variabili casuali x_n . Quando tali distribuzioni non si conoscono, che è poi la situazione tipica nelle applicazioni dell'elaborazione dei segnali, si assume, spesso legittimamente, una legge di probabilità gaussiana. Da quest'ipotesi sulla distribuzione probabilistica delle variabili casuali x_n è spesso possibile ricavare limiti di confidenza approssimati per le stime della media, della varianza, etc. Molte volte, però, ci si accontenta delle espressioni della polarizzazione e della varianza degli stimatori. Inoltre, per orientare l'applicazione delle tecniche di elaborazione numerica dei segnali alla stima delle medie di segnali aleatori, bastano anche espressioni approssimate, che evidenzino la dipendenza della polarizzazione e della varianza dalla lunghezza della sequenza campione. Nei prossimi paragrafi ci occuperemo quindi di ricavare delle espressioni per la polarizzazione e la varianza di vari stimatori dell'autocovarianza e dello spettro di potenza di un segnale aleatorio stazionario, e useremo queste espressioni per approfondire i problemi incontrati nel calcolo delle stime.

11.2 STIME DELL'AUTOCOVARIANZA

Per studiare gli stimatori della sequenza di autocovarianza di un processo aleatorio si possono usare i concetti introdotti nel paragrafo precedente. Consideriamo ancora un processo aleatorio stazionario $\{x_n\}$, $-\infty < n < \infty$, e per comodità assumiamo che sia a media nulla, cioè

$$m_x = E[x_n] = 0, \quad \text{qualsiasi } n$$

La sequenza di autocovarianza è allora

$$\gamma_{xx}(m) = E[x_n x_{n+m}^*]$$

che è anche uguale alla sequenza di autocorrelazione, $\phi_{xx}(m)$. Ci riferiremo perciò d'ora innanzi a stime della sequenza di autocorrelazione,

tenendo presente che l'autocorrelazione e l'autocovarianza sono identiche per un processo a media nulla. Assumiamo inoltre che sia

$$\gamma_{xx}(m) = \langle x(n)x^*(n+m) \rangle \quad (11.16)$$

per tutte le sequenze campione $x(n)$. Possiamo riscrivere la (11.16) come

$$\gamma_{xx}(m) = \langle g_m(n) \rangle$$

dove è $g_m(n) = x(n)x^*(n+m)$. Di conseguenza, si può vedere il problema di stimare l'autocovarianza di un processo a media nulla come quello di stimare la media di $g_m(n)$. Dati N valori consecutivi della sequenza $x(n)$, abbiamo a disposizione $(N-m)$ campioni consecutivi di $g_m(n)$ in base ai quali stimarne la media. Applicando la definizione di media campione introdotta nel paragrafo precedente, otteniamo come stima della sequenza di autocorrelazione

$$c'_{xx}(m) = \frac{1}{N-|m|} \sum_{n=0}^{N-|m|-1} x(n)x^*(n+m) \quad (11.17)$$

dove è $|m| < N$. Se la sequenza $g_m(n)$ è gaussiana, allora l'espressione (11.17) rappresenta la stima a massima verosimiglianza della sequenza di autocorrelazione. In generale, invece, il procedimento formale che serve a ricavare la stima a massima verosimiglianza conduce a un insieme di equazioni che non possono essere risolte neanche se la legge di probabilità di $g_m(n)$ è nota. Tuttavia l'espressione (11.17), anche se può non essere formalmente ottima, rappresenta una scelta ragionevole per uno stimatore della sequenza di autocorrelazione. Si vede facilmente che $c'_{xx}(m)$ è una stima non polarizzata di $\phi_{xx}(m)$ essendo $E[x(n)x^*(n+m)] = \phi_{xx}(m)$. La varianza di $c'_{xx}(m)$ si può ricavare come nel par. 11.1; i passaggi matematici sono però pesanti e non li riportiamo qui. Un'espressione approssimata della varianza, fornita da Jenkins e Watts [5], è

$$\text{var}[c'_{xx}(m)] \cong \frac{N}{[N-|m|]^2} \sum_{r=-\infty}^{\infty} [\phi_{xx}^2(r) + \phi_{xx}(r+m)\phi_{xx}(r-m)] \quad (11.18)$$

Questa espressione vale per N molto maggiore di m ; comunque, la varianza di $c'_{xx}(m)$ è in generale proporzionale a $1/N$ come nella (11.18). Essendo la polarizzazione nulla e

$$\lim_{N \rightarrow \infty} \{\text{var}[c'_{xx}(m)]\} \rightarrow 0$$

$c'_{xx}(m)$ è una stima consistente di $\phi_{xx}(m)$.

Un altro stimatore della sequenza di autocorrelazione è

$$c_{xx}(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n)x(n+m) \quad (11.19)$$

Questa stima differisce dalla stima $c'_{xx}(m)$ definita nella (11.17) solo per il fattore moltiplicativo che precede la sommatoria. In effetti, confrontando le (11.17) e (11.19), si vede che è

$$c_{xx}(m) = \frac{N-|m|}{N} c'_{xx}(m) \quad (11.20)$$

Poiché il valore atteso di $c'_{xx}(m)$ è $\phi_{xx}(m)$, il valore atteso di $c_{xx}(m)$ è

$$E[c_{xx}(m)] = \frac{N-|m|}{N} \phi_{xx}(m) \quad (11.21)$$

Di conseguenza, $c_{xx}(m)$ è una stima polarizzata della sequenza di autocorrelazione, anche se è asintoticamente non polarizzata. In particolare, la polarizzazione della stima $c_{xx}(m)$ vale

$$\text{polarizzazione} = \phi_{xx}(m) \left[\frac{m}{N} \right] \quad (11.22)$$

Dalla relazione (11.20) segue che la varianza di $c_{xx}(m)$ è pari a $[(N-|m|)N]^2$ volte la varianza di $c'_{xx}(m)$, in modo che per N grande rispetto ad m si ha

$$\text{var}[c_{xx}(m)] \cong \frac{1}{N} \sum_{r=-\infty}^{\infty} [\phi_{xx}^2(r) + \phi_{xx}(r+m)\phi_{xx}(r-m)] \quad (11.23)$$

Quando il valore di m tende a N , la varianza della stima $c'_{xx}(m)$ cresce notevolmente. Questa è una conseguenza del fatto che la stima $c'_{xx}(m)$ è ottenuta calcolando la media campione della sequenza $g_m(n)$. Se m è dell'ordine di N , allora restano pochi punti a disposizione da usare nel calcolo della media campione di $g_m(n)$ ed è per questo motivo che, quando m si avvicina alla lunghezza della sequenza, la varianza della stima $c'_{xx}(m)$ diventa grande. Di conseguenza, non si ha una stima utile. Se invece consideriamo lo stimatore polarizzato $c_{xx}(m)$, osserviamo che la sua varianza non ha la stessa tendenza a crescere quando m è dell'ordine della lunghezza della sequenza. Però, quando m tende a N , la polarizzazione tende a $\phi_{xx}(m)$, cioè il valor medio della stima tende a zero. Poiché la polarizzazione è grande quanto la funzione che stiamo stimando, neppure questa può essere considerata una buona stima quando m è dell'ordine di N .

Le conclusioni che abbiamo tratte si basano su un esame della polarizzazione e della varianza quando il valore m del ritardo cresce e la lunghezza del tratto di sequenza analizzato è fissa. Possiamo esaminare il comportamento della polarizzazione e della varianza anche da un altro punto di vista, cioè tenendo costante il valore del ritardo m e facendo crescere la lunghezza N della sequenza. Per m fissato si vede dall'espressione (11.18) che la varianza della stima non polarizzata $c'_{xx}(m)$ diminuisce al crescere di N . Per quanto riguarda la stima polarizzata, risulta dalle (11.22) e (11.23) che al crescere di N diminuiscono sia la polarizzazione che la varianza. Jenkins e Watts [5] fanno l'ipotesi che in molti casi l'errore quadratico medio relativo allo stimatore polarizzato sia minore di quello relativo allo stimatore non polarizzato. Se questa congettura è valida, essa fornisce una motivazione per preferire lo stimatore polarizzato $c_{xx}(m)$. Comunque, entrambi gli stimatori sono asintoticamente non polarizzati, per cui ci si può aspettare che in generale la stima della sequenza di autocorrelazione migliori usando per essa un maggior numero di campioni.

11.3 IL PERIODOGRAMMA COME STIMA DELLO SPETTRO DI POTENZA

Nel paragrafo precedente abbiamo introdotto due plausibili stimatori della sequenza di autocovarianza e abbiamo visto che essi forniscono stime consistenti e asintoticamente non polarizzate. Saremmo quindi tentati di concludere che le trasformate di Fourier di tali stime della sequenza di autocovarianza forniscono buone stime della densità spettrale di potenza. Sfortunatamente, questo non è vero. In particolare, dimostreremo che le trasformate di Fourier di stime consistenti della covarianza non sono stime consistenti dello spettro di potenza poiché la varianza corrispondente non tende a zero al crescere della lunghezza della sequenza, N . Vedremo però che è possibile ricavare una buona stima dello spettro di potenza « smussando » la trasformata di Fourier della stima della covarianza.

In generale, le espressioni esatte della varianza delle stime dello spettro sono troppo complicate. Pertanto, conviene orientare la valutazione delle stime dello spettro verso espressioni approssimate di facile interpretazione. Per questo molte delle espressioni che saranno ricavate nel seguito sono soltanto approssimate: comunque, verrà fatto notare ogni volta in cosa consiste l'approssimazione.

11.3.1 Definizione del periodogramma

Consideriamo il caso in cui si prenda come stima della densità spettrale di potenza la trasformata di Fourier della stima polarizzata dell'autocorrelazione $c_{xx}(m)$, cioè

$$I_N(\omega) = \sum_{m=-(N-1)}^{N-1} c_{xx}(m) e^{-j\omega m} \quad (11.24)$$

Poiché la trasformata di Fourier della sequenza reale e di lunghezza finita $x(n)$, $0 \leq n \leq N-1$ è

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n) e^{-j\omega n}$$

si può dimostrare (v. probl. 1 di questo capitolo) che è

$$I_N(\omega) = \frac{1}{N} |X(e^{j\omega})|^2 \quad (11.25)$$

La stima dello spettro $I_N(\omega)$ è spesso chiamata *periodogramma*.

Come nei casi precedenti, è interessante trovare la polarizzazione e la varianza del periodogramma usato come stima dello spettro di potenza. Il valore atteso di $I_N(\omega)$ è

$$E[I_N(\omega)] = \sum_{m=-(N-1)}^{N-1} E[c_{xx}(m)] e^{-j\omega m} \quad (11.26)$$

Poiché abbiamo dimostrato che per un processo a media nulla è

$$E[c_{xx}(m)] = \frac{N-|m|}{N} \phi_{xx}(m), \quad |m| < N$$

allora si ha

$$E[I_N(\omega)] = \sum_{m=-(N-1)}^{N-1} \left(\frac{N-|m|}{N} \right) \phi_{xx}(m) e^{-j\omega m} \quad (11.27)$$

Quindi, a causa dei limiti finiti nella sommatoria e del fattore $(N-|m|)/N$, $E[I_N(\omega)]$ non coincide con la trasformata di Fourier di $\phi_{xx}(m)$, e perciò il periodogramma è una stima polarizzata dello spettro di potenza $P_{xx}(\omega)$.

In alternativa, consideriamo la trasformata di Fourier della stima $c'_{xx}(m)$, vale a dire

$$P_N(\omega) = \sum_{m=-(N-1)}^{N-1} c'_{xx}(m) e^{-j\omega m} \quad (11.28)$$

Il valore atteso di $P_N(\omega)$ è

$$\begin{aligned} E[P_N(\omega)] &= \sum_{m=-(N-1)}^{N-1} E[c'_{xx}(m)] e^{-j\omega m} \\ &= \sum_{m=-(N-1)}^{N-1} \phi_{xx}(m) e^{-j\omega m} \end{aligned} \quad (11.29)$$

Di nuovo, a causa dei limiti finiti della sommatoria, questa è una stima polarizzata di $P_{xx}(\omega)$, anche se $c'_{xx}(m)$ è una stima non polarizzata di $\phi_{xx}(m)$.

Possiamo interpretare le formule (11.27) e (11.29) come trasformate di Fourier di sequenze di autocorrelazione pesate con finestre. Nel caso della (11.27) la finestra è quella triangolare

$$w_B(m) = \begin{cases} \frac{N-|m|}{N}, & |m| < N \\ 0, & \text{altrove} \end{cases} \quad (11.30)$$

che nel cap. 5 abbiamo chiamato finestra di Bartlett. Per la (11.29) la finestra è rettangolare, cioè

$$w_R(n) = \begin{cases} 1, & |m| < N \\ 0, & \text{altrove} \end{cases} \quad (11.31)$$

Usando i concetti introdotti nel cap. 5 si può vedere che le relazioni (11.27) e (11.29) sono interpretabili nel dominio della frequenza come le convoluzioni

$$E[I_N(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\theta) W_B(e^{j(\omega-\theta)}) d\theta \quad (11.32)$$

e

$$E[P_N(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\theta) W_R(e^{j(\omega-\theta)}) d\theta \quad (11.33)$$

dove

$$W_B(e^{j\omega}) = \frac{1}{N} \left(\frac{\sin [\omega N/2]}{\sin [\omega/2]} \right)^2 \quad (11.34)$$

e

$$W_R(e^{j\omega}) = \frac{\sin [\omega(2N-1)/2]}{\sin [\omega/2]} \quad (11.35)$$

sono le trasformate di Fourier, rispettivamente, della finestra di Bartlett e di quella rettangolare.

11.3.2 Varianza del periodogramma

Per ottenere un'espressione per la varianza del periodogramma, è opportuno innanzitutto fare l'ipotesi che la sequenza $x(n)$, $0 \leq n \leq N-1$, sia una sequenza campione di un processo reale, bianco e gaussiano. Il periodogramma $I_N(\omega)$ si può riscrivere come

$$\begin{aligned} I_N(\omega) &= \frac{1}{N} |X(e^{j\omega})|^2 \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \sum_{m=0}^{N-1} x(l)x(m)e^{j\omega m}e^{-j\omega l} \end{aligned}$$

Per calcolare la covarianza di $I_N(\omega)$ a due frequenze ω_1 e ω_2 consideriamo dapprima

$$E[I_N(\omega_1)I_N(\omega_2)] = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} E[x(k)x(l)x(m)x(n)]e^{j[\omega_1(k-l)+\omega_2(m-n)]} \quad (11.36)$$

Per ottenere un risultato utile occorre semplificare la (11.36). In generale, non è possibile ottenere un risultato molto semplice neppure quando $x(n)$ è bianco, in quanto il fatto che sia $E[x(n)x(n+m)] = \sigma_x^2 \delta(m)$ non assicura un'espressione semplice di $E[x(k)x(l)x(m)x(n)]$ per tutte le combinazioni di k, l, m ed n . Si può dimostrare tuttavia [7] che, nel caso di un processo bianco gaussiano, risulta

$$\begin{aligned} E[x(k)x(l)x(m)x(n)] &= E[x(k)x(l)]E[x(m)x(n)] \\ &\quad + E[x(k)x(m)]E[x(l)x(n)] \\ &\quad + E[x(k)x(n)]E[x(l)x(m)] \end{aligned}$$

Si ha allora

$$E[x(k)x(l)x(m)x(n)] = \begin{cases} \sigma_x^4, & k=l \text{ e } m=n \\ & \text{o } k=m \text{ e } l=n \\ & \text{o } k=n \text{ e } l=m \\ 0, & \text{altrimenti} \end{cases} \quad (11.37)$$

Quando il processo non è gaussiano, il risultato non è generalmente altrettanto semplice. Il nostro scopo è però quello di illustrare i problemi della stima dello spettro, piuttosto che ricavare una formula generale valida

sempre ma di difficile interpretazione. Sostituendo dunque la (11.37) nella (11.36) otteniamo

$$E[I_N(\omega_1)I_N(\omega_2)] = \frac{\sigma_x^4}{N^2} \left\{ N^2 + \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} e^{j(m-n)(\omega_1+\omega_2)} + \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} e^{j(n-m)(\omega_1-\omega_2)} \right\}$$

ovvero

$$\begin{aligned} E[I_N(\omega_1)I_N(\omega_2)] &= \sigma_x^4 \left\{ 1 + \left(\frac{\sin [(\omega_1 + \omega_2)N/2]}{N \sin [(\omega_1 + \omega_2)/2]} \right)^2 \right. \\ &\quad \left. + \left(\frac{\sin [(\omega_1 - \omega_2)N/2]}{N \sin [(\omega_1 - \omega_2)/2]} \right)^2 \right\} \quad (11.38) \end{aligned}$$

(Se il segnale non è gaussiano, l'espressione (11.38) contiene termini aggiuntivi che sono proporzionali a $1/N$ [4,8]). La covarianza del periodogramma è

$$\text{cov} [I_N(\omega_1), I_N(\omega_2)] = E[I_N(\omega_1)I_N(\omega_2)] - E[I_N(\omega_1)]E[I_N(\omega_2)] \quad (11.39)$$

Essendo $E[I_N(\omega_1)] = E[I_N(\omega_2)] = \sigma_x^2$, si ricava dalle (11.38) e (11.39)

$$\begin{aligned} \text{cov} [I_N(\omega_1), I_N(\omega_2)] &= \sigma_x^4 \left\{ \left(\frac{\sin [(\omega_1 + \omega_2)N/2]}{N \sin [(\omega_1 + \omega_2)/2]} \right)^2 \right. \\ &\quad \left. + \left(\frac{\sin [(\omega_1 - \omega_2)N/2]}{N \sin [(\omega_1 - \omega_2)/2]} \right)^2 \right\} \quad (11.40) \end{aligned}$$

Da questa espressione si possono trarre molte conclusioni interessanti sul periodogramma. La varianza della stima dello spettro ad una particolare frequenza $\omega = \omega_1 = \omega_2$ è

$$\text{var} [I_N(\omega)] = \text{cov} [I_N(\omega), I_N(\omega)] = \sigma_x^4 \left\{ 1 + \left(\frac{\sin [\omega N]}{N \sin \omega} \right)^2 \right\} \quad (11.41)$$

Chiaramente, la varianza di $I_N(\omega)$ non tende a zero quando N tende all'infinito. Perciò il periodogramma non è una stima consistente: in effetti, $\text{var} [I_N(\omega)]$ è dell'ordine di σ_x^4 indipendentemente da N .

Dalla relazione (11.40) si vede anche che alle frequenze $\omega_1 = 2\pi k/N$ e $\omega_2 = 2\pi l/N$, con k ed l interi, risulta

$$\text{cov} [I_N(\omega_1), I_N(\omega_2)] = \sigma_x^4 \left\{ \left(\frac{\sin [\pi(k+l)]}{N \sin [\pi(k+l)/N]} \right)^2 + \left(\frac{\sin [\pi(k-l)]}{N \sin [\pi(k-l)/N]} \right)^2 \right\} \quad (11.42)$$

che è uguale a zero per $k \neq l$. Quindi i valori del periodogramma che distano in frequenza di multipli interi di $2\pi/N$ sono scorrelati. Quando N aumenta, questi campioni in frequenza con covarianza nulla si avvicinano tra loro. Avendo assunto che il processo è bianco, è ragionevole attendersi che una buona stima dello spettro di potenza tenda a un valore costante al crescere di N . Invece, in seguito al fatto che al crescere di N la varianza

del periodogramma tende a una costante diversa da zero e che la distanza tra i campioni dello spettro aventi covarianza nulla diminuisce, la rapidità di variazione del periodogramma aumenta all'aumentare della lunghezza della sequenza. Questo comportamento è illustrato nella fig. 11.3, che rappresenta il periodogramma per lunghezze di $N = 14, 51, 135$ e 452 campioni.

11.3.3 Espressioni generali per la varianza

Tutta la discussione precedente era relativa alla stima dello spettro di un rumore bianco. Se consideriamo un processo che non è bianco pur essendo gaussiano, l'analisi diventa notevolmente più difficile. In questo caso più generale, è utile affrontare il problema del calcolo della covarianza tra campioni dello spettro in maniera euristica ed arrivare a una espressione approssimata. Questo è quanto faremo noi, mentre una derivazione più rigorosa è fornita da Jenkins e Watts [5]. I risultati ricavati qui possono essere ottenuti dai loro introducendovi alcune approssimazioni. Il punto di partenza per l'impostazione euristica del problema è il fatto che una sequenza aleatoria non bianca (ovvero colorata) può essere generata facendo passare del rumore bianco attraverso un sistema lineare. La densità spettrale di potenza del rumore d'uscita è il prodotto della densità spettrale di potenza dell'ingresso e del modulo al quadrato della risposta in frequenza del sistema. Consideriamo ora una realizzazione finita di un rumore non bianco, la cui lunghezza indichiamo con N . Naturalmente, non è del tutto vero che si può ottenere un segmento di rumore non bianco filtrando con un sistema lineare un segmento di rumore bianco, a causa degli effetti di transitorio all'inizio e alla fine del segmento. Tuttavia, se la durata della sequenza è lunga rispetto a quella della risposta all'impulso del filtro, sembra per lo meno ragionevole approssimare una realizzazione di rumore non bianco in questo modo. Consideriamo adesso un processo gaussiano non bianco con densità spettrale di potenza $P_{xx}(\omega)$. Indichiamo con $x_N(n)$ una realizzazione lunga N del rumore non bianco e con $w_N(n)$ una realizzazione lunga N di rumore bianco con varianza unitaria. La nostra approssimazione è allora che $x_N(n)$ sia il risultato del filtraggio di $w_N(n)$ con un sistema lineare la cui risposta in frequenza abbia modulo quadrato pari a $P_{xx}(\omega)$. Se $I_N(\omega)$ indica il periodogramma del rumore colorato e $I_N^w(\omega)$ il periodogramma del rumore bianco, si ha

$$I_N(\omega) = \frac{1}{N} |X_N(e^{j\omega})|^2$$

$$I_N^w(\omega) = \frac{1}{N} |W_N(e^{j\omega})|^2$$

ed essendo

$$|X_N(e^{j\omega})|^2 \cong P_{xx}(\omega) |W_N(e^{j\omega})|^2$$

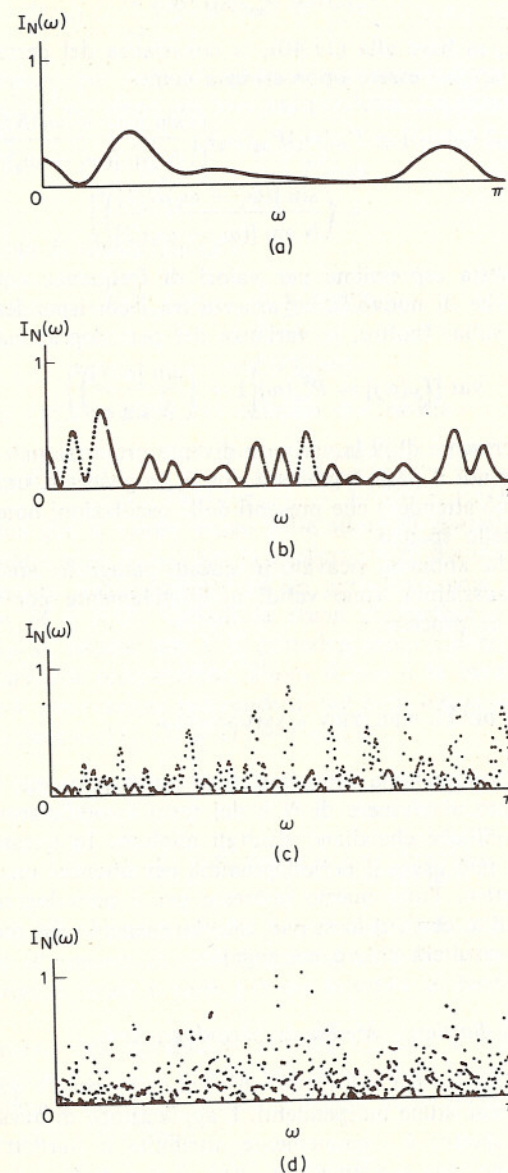


Fig. 11.3 Periodogrammi relativi a sequenze di lunghezza $N =$ (a) 14, (b) 51, (c) 135 e (d) 452, che presentano oscillazioni crescenti al crescere di N .

ne consegue che è

$$I_N(\omega) \cong P_{xx}(\omega) I_N^w(\omega)$$

Di conseguenza, in base alla (11.40), la covarianza del periodogramma a frequenze diverse può essere approssimata come

$$\begin{aligned} \text{cov}[I_N(\omega_1), I_N(\omega_2)] &\cong P_{xx}(\omega_1) P_{xx}(\omega_2) \left\{ \left(\frac{\sin[(\omega_1 + \omega_2)N/2]}{N \sin[(\omega_1 + \omega_2)/2]} \right)^2 \right. \\ &\quad \left. + \left(\frac{\sin[(\omega_1 - \omega_2)N/2]}{N \sin[(\omega_1 - \omega_2)/2]} \right)^2 \right\} \end{aligned} \quad (11.43)$$

Se si valuta questa espressione per valori di frequenza equispaziati di $2\pi/N$, si vede che di nuovo la covarianza tra i corrispondenti campioni in frequenza è nulla. Inoltre, la varianza del periodogramma è

$$\text{var}[I_N(\omega)] = P_{xx}^2(\omega) \left\{ 1 + \left(\frac{\sin[\omega N]}{N \sin \omega} \right)^2 \right\} \quad (11.44)$$

in modo che al crescere di N la varianza diventa proporzionale al quadrato dello spettro. Quindi il periodogramma non è, in generale, una stima consistente e ci si può attendere che presenti delle oscillazioni notevoli attorno al valore vero dello spettro.

I risultati che abbiamo ricavato in questo paragrafo, anche se basati sull'ipotesi di gaussianità, sono validi qualitativamente per una gamma piuttosto ampia di processi.

11.4 STIMATORI DELLO SPETTRO « SMUSSATI »

Poiché il periodogramma non è una stima consistente dello spettro ed il suo comportamento al crescere di N è del tutto insoddisfacente, occorre studiare delle modifiche che diano risultati migliori. In questo paragrafo vedremo come si può usare il periodogramma per ottenere una stima consistente dello spettro. Tutto questo interesse per il periodogramma è giustificato dalla facilità con cui lo si può calcolare usando dei metodi basati sulla FFT, come risulterà chiaro nel seguito.

11.4.1 Metodo di Bartlett - Media di periodogrammi

Un metodo classico per ridurre la varianza delle stime è quello di fare la media di numerose stime indipendenti. L'applicazione di questo concetto alla stima dello spettro è comunemente attribuita a Bartlett. Seguendo questa impostazione, una sequenza dati $x(n)$, $0 \leq n \leq N-1$, viene suddivisa in K segmenti di M campioni ciascuno (per cui è $N = KM$); si costruiscono cioè le sottosequenze

$$x^{(i)}(n) = x(n + iM - M), \quad 0 \leq n \leq M-1, \quad 1 \leq i \leq K \quad (11.45)$$

e si calcolano i K periodogrammi

$$I_M^{(i)}(\omega) = \frac{1}{M} \left| \sum_{n=0}^{M-1} x^{(i)}(n) e^{-j\omega n} \right|^2, \quad 1 \leq i \leq K \quad (11.46)$$

Se $\phi_{xx}(m)$ è piccolo per $m > M$, allora è ragionevole fare l'ipotesi che i periodogrammi $I_M^{(i)}(\omega)$ siano tra loro indipendenti. La stima dello spettro è definita come

$$B_{xx}(\omega) = \frac{1}{K} \sum_{i=1}^K I_M^{(i)}(\omega) \quad (11.47)$$

e il valore atteso di questa stima è

$$\begin{aligned} E[B_{xx}(\omega)] &= \frac{1}{K} \sum_{i=1}^K E[I_M^{(i)}(\omega)] \\ &= E[I_M^{(i)}(\omega)] \end{aligned}$$

Dalle relazioni (11.32) e (11.34) notiamo che risulta

$$E[B_{xx}(\omega)] = E[I_M^{(i)}(\omega)] = \frac{1}{2\pi M} \int_{-\pi}^{\pi} P_{xx}(\theta) \left(\frac{\sin[(\omega - \theta)M/2]}{\sin[(\omega - \theta)/2]} \right)^2 d\theta \quad (11.48)$$

Questo significa che il valore atteso della stima di Bartlett è la convoluzione dello spettro vero $P_{xx}(\omega)$ con la trasformata di Fourier della funzione finestra triangolare corrispondente ad un periodogramma calcolato su M campioni, con $M = N/K$. Quindi la stima di Bartlett è uno stimatore polarizzato. Se si assume che i K periodogrammi mediati nella (11.47) siano statisticamente indipendenti, allora $B_{xx}(\omega)$ è la media campione di un insieme di K osservazioni indipendenti del periodogramma $I_M(\omega)$. Segue quindi dalle relazioni (11.11) e (11.44) che è

$$\begin{aligned} \text{var}[B_{xx}(\omega)] &= \frac{1}{K} \text{var}[I_M(\omega)] \\ &\cong \frac{1}{K} P_{xx}^2(\omega) \left\{ 1 + \left(\frac{\sin[\omega M]}{M \sin \omega} \right)^2 \right\} \end{aligned} \quad (11.49)$$

È chiaro da questa espressione che la varianza di $B_{xx}(\omega)$ è inversamente proporzionale al numero di periodogrammi mediati, e quando K diventa grande la varianza tende a zero, per cui la stima di Bartlett è una stima consistente.

Il confronto dell'espressione (11.48) relativa a $E[B_{xx}(\omega)]$ con la (11.32) relativa a $E[I_N(\omega)]$ mostra che in entrambi i casi il valore atteso della stima appare in forma di convoluzione dello spettro vero con una « finestra spettrale » del tipo

$$W_B(e^{j\omega}) = \frac{1}{N'} \left(\frac{\sin[\omega N']}{\sin \omega} \right)^2$$

dove è $N' = N$ per il periodogramma e $N' = M = N/K$ per la stima di Bartlett. La polarizzazione di $B_{xx}(\omega)$ è maggiore di quella di $I_N(\omega)$, a causa

della maggior larghezza del lobo principale della finestra spettrale. La polarizzazione può dunque essere interpretata in termini dei suoi effetti sulla risoluzione dello spettro. Per una lunghezza fissata della sequenza da analizzare, la varianza diminuisce al crescere del numero dei periodogrammi, ma diminuisce anche M e quindi la risoluzione dello spettro. Occorre perciò raggiungere un compromesso, nel metodo di Bartlett, tra polarizzazione o risoluzione dello spettro da un lato e varianza della stima dall'altro. La scelta effettiva di M ed N per la stima dello spettro in una situazione reale sarà orientata in generale dalle conoscenze a priori sul segnale da analizzare. Per esempio, se sappiamo che lo spettro ha un picco molto stretto, e se è importante risolverlo, dobbiamo scegliere M abbastanza grande per ottenere la risoluzione in frequenza desiderata. Dall'espressione della varianza si può risalire alla lunghezza della sequenza $N = KM$ che dia una varianza accettabile della stima.

11.4.2 Metodo delle finestre

Abbiamo visto che la varianza della stima di Bartlett dello spettro può essere ridotta a spese di un aumento della polarizzazione e di una diminuzione della risoluzione. Nel metodo di Bartlett la risoluzione dello spettro diminuisce perché si usano sequenze più corte. Un altro metodo consiste nello smussare il periodogramma facendone la convoluzione con una opportuna finestra spettrale [2]. In altre parole, il periodogramma smussato $S_{xx}(\omega)$ è

$$S_{xx}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} I_N(\theta) W(e^{j(\omega-\theta)}) d\theta \quad (11.50)$$

dove $W(e^{j\omega})$ è la finestra spettrale. Poiché il periodogramma è la trasformata di Fourier di $c_{xx}(m)$, allora $S_{xx}(\omega)$ risulta essere la trasformata di Fourier del prodotto di $c_{xx}(m)$ e della trasformata di Fourier inversa di $W(e^{j\omega})$. Perciò se

$$w(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\omega}) e^{j\omega m} d\omega$$

è una sequenza di durata finita lunga $2M - 1$, allora si ha

$$S_{xx}(\omega) = \sum_{m=-(M-1)}^{M-1} c_{xx}(m) w(m) e^{-j\omega m} \quad (11.51)$$

Affinché $S_{xx}(\omega)$ sia una funzione reale e pari quando la sequenza dei dati $x(n)$ è reale, la finestra $w(m)$ deve essere una sequenza pari. Inoltre, ricordiamo che lo spettro di potenza è una funzione non negativa della frequenza, per cui è ragionevole richiedere che anche $S_{xx}(\omega)$ sia non negativo. Si noti che sia il periodogramma che la stima di Bartlett sono funzioni non negative della frequenza. Dall'espressione (11.50) è però chiaro che una

condizione sufficiente (anche se certamente non necessaria) perché $S_{xx}(\omega)$ sia non negativo è che risulti

$$W(e^{j\omega}) \geq 0, \quad -\pi \leq \omega \leq \pi$$

Questa condizione è verificata per la finestra di Bartlett o triangolare ma non vale, ad esempio, per le finestre di Hamming o di Hanning. Perciò queste ultime sequenze finestra, anche se assicurano una migliore risoluzione in frequenza e lobi laterali più bassi, possono dar luogo a stime dello spettro negative in qualche intervallo di frequenza.

Si vede facilmente che il valore atteso dell'espressione (11.51) è

$$E[S_{xx}(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} E[I_N(\theta)] W(e^{j(\omega-\theta)}) d\theta \quad (11.52)$$

Poiché si ha dalla (11.32)

$$E[I_N(\theta)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\phi) W_B(e^{j(\theta-\phi)}) d\phi$$

vediamo che $E[S_{xx}(\omega)]$ è la convoluzione nel dominio della frequenza di $W_B(e^{j\omega})$ e $W(e^{j\omega})$ con $P_{xx}(\omega)$. Quindi, $E[S_{xx}(\omega)]$ è la trasformata di Fourier di $\phi_{xx}(m)$ moltiplicata per il prodotto della finestra triangolare $w_B(m)$ con $w(m)$; si ha cioè

$$E[S_{xx}(\omega)] = \sum_{m=-(M-1)}^{M-1} c_{xx}(m) w_B(m) w(m) e^{-j\omega m} \quad (11.53)$$

con

$$w_B(m) = 1 - \frac{|m|}{N}, \quad |m| < N$$

Se M è piccolo rispetto a N , allora $W(e^{j\omega})$ è la larga rispetto a $W_B(e^{j\omega})$ e la (11.52) si può riscrivere approssimativamente come

$$E[S_{xx}(\omega)] \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\theta) W(e^{j(\omega-\theta)}) d\theta \quad (11.54)$$

Dall'espressione (11.52) o dalla (11.54) si vede che un aumento della larghezza della finestra spettrale ha l'effetto di smussare ulteriormente lo spettro e di ridurre la risoluzione in frequenza della stima.

Per studiare gli effetti della finestra sulla varianza della stima dello spettro, si può calcolare, analogamente a quanto si è fatto nel par. 11.3.2, la covarianza del periodogramma smussato. La covarianza alle due frequenze ω_1 e ω_2 vale

$$\text{cov}[S_{xx}(\omega_1), S_{xx}(\omega_2)] = E[(S_{xx}(\omega_1) - E[S_{xx}(\omega_1)])(S_{xx}(\omega_2) - E[S_{xx}(\omega_2)])]$$

Dalle (11.50) e (11.52) segue che è

$$S_{xx}(\omega) - E[S_{xx}(\omega)] \cong \frac{1}{2\pi} \int_{-\pi}^{\pi} (I_N(\theta) - E[I_N(\theta)]) W(e^{j(\omega-\theta)}) d\theta$$

per cui risulta

$$\text{cov}[S_{xx}(\omega_1), S_{xx}(\omega_2)] \cong \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} W(e^{j(\omega_1-\theta)}) W(e^{j(\omega_2-\phi)}) \times \text{cov}[I_N(\theta), I_N(\phi)] d\theta d\phi \quad (11.55)$$

Si ha però dalla relazione (11.43)

$$\text{cov}[I_N(\theta), I_N(\phi)] \cong P_{xx}(\theta) P_{xx}(\phi) \left[\left(\frac{\sin[(\theta + \phi)N/2]}{N \sin[(\theta + \phi)/2]} \right)^2 + \left(\frac{\sin[(\theta - \phi)N/2]}{N \sin[(\theta - \phi)/2]} \right)^2 \right]$$

Se assumiamo che i termini

$$\left(\frac{\sin[(\theta + \phi)N/2]}{N \sin[(\theta + \phi)/2]} \right)^2$$

e

$$\left(\frac{\sin[(\theta - \phi)N/2]}{N \sin[(\theta - \phi)/2]} \right)^2$$

siano a banda stretta rispetto alle variazioni di $P_{xx}(\omega)$ e $W(e^{j\omega})$, e che siano molto concentrati intorno a $\theta = -\phi$ e $\theta = \phi$ rispettivamente (cioè che N sia grande), allora si ottiene, approssimando prima l'integrale su θ^1 ,

$$\text{cov}[S_{xx}(\omega_1), S_{xx}(\omega_2)] \cong \frac{1}{2\pi N} \int_{-\pi}^{\pi} P_{xx}^2(\phi) W(e^{j(\omega_2-\phi)}) [W(e^{j(\omega_1+\phi)}) + W(e^{j(\omega_1-\phi)})] d\phi \quad (11.56)$$

Se assumiamo inoltre che la finestra spettrale sia abbastanza stretta da poter trascurare il termine $W(e^{j(\omega_1+\phi)}) W(e^{j(\omega_2-\phi)})$ l'espressione precedente diventa

$$\text{cov}[S_{xx}(\omega_1), S_{xx}(\omega_2)] \cong \frac{1}{2\pi N} \int_{-\pi}^{\pi} P_{xx}^2(\phi) W(e^{j(\omega_1-\phi)}) W(e^{j(\omega_2-\phi)}) d\phi \quad (11.57)$$

Da questa formula è chiaro che quando la larghezza della finestra spettrale $W(e^{j\omega})$ aumenta, in modo che vi è maggiore sovrapposizione tra $W(e^{j(\omega_1-\phi)})$ e $W(e^{j(\omega_2-\phi)})$, aumenta anche la covarianza tra stime fatte a frequenze diverse.

Per ricavare la varianza della stima dello spettro $S_{xx}(\omega)$, basta valutare la (11.57) per $\omega = \omega_1 = \omega_2$ e si ottiene

$$\text{var}[S_{xx}(\omega)] \cong \frac{1}{2\pi N} \int_{-\pi}^{\pi} P_{xx}^2(\phi) W^2(e^{j(\omega-\phi)}) d\phi \quad (11.58)$$

Assumiamo adesso che $W(e^{j\omega})$ sia stretta rispetto alle variazioni di $P_{xx}(\omega)$, facciamo cioè l'ipotesi di aver potuto scegliere la lunghezza della finestra

¹ Usiamo qui il fatto che è

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\sin[\theta N/2]}{N \sin[\theta/2]} \right)^2 d\theta = \frac{1}{N}$$

(v. probl. 3 di questo capitolo)

$w(m)$ abbastanza grande per avere la necessaria risoluzione in frequenza. Si può allora approssimare ulteriormente l'espressione (11.58) con la

$$\text{var}[S_{xx}(\omega)] \cong \frac{1}{N} P_{xx}^2(\omega) \frac{1}{2\pi} \int_{-\pi}^{\pi} W^2(e^{j\phi}) d\phi \quad (11.59a)$$

Notando che il teorema di Parseval, sotto l'ipotesi di $w(m)$ simmetrica, permette di scrivere

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} W^2(e^{j\phi}) d\phi = \sum_{m=-(M-1)}^{M-1} w^2(m)$$

otteniamo l'espressione più comoda

$$\text{var}[S_{xx}(\omega)] \cong \left(\frac{1}{N} \sum_{m=-(M-1)}^{M-1} w^2(m) \right) P_{xx}^2(\omega) \quad (11.59b)$$

Le formule (11.54) e (11.59) sono delle espressioni approssimate per la media e la varianza della stima dello spettro $S_{xx}(\omega)$. Esse valgono sotto l'ipotesi che la lunghezza $(2M-1)$ della finestra $w(m)$ applicata alla stima $c_{xx}(m)$ sia tale da poter considerare $W(e^{j\omega})$ a banda stretta rispetto alle variazioni dello spettro $P_{xx}(\omega)$ e contemporaneamente a banda larga rispetto a $(\sin[\omega N/2]/\sin[\omega/2])^2$. Per valutare i miglioramenti apportati dall'uso della finestra si possono confrontare queste espressioni con quelle corrispondenti relative al periodogramma.

Dalla (11.27) si vede che il periodogramma è asintoticamente non polarizzato, cioè

$$\lim_{N \rightarrow \infty} E[I_N(\omega)] = P_{xx}(\omega)$$

In base alla (11.54) si vede che quando la lunghezza N della sequenza diventa grande, si può anche rendere lunga la finestra, in modo che $W(e^{j\omega})$ sia stretta rispetto alle variazioni di $P_{xx}(\omega)$; questo implica allora che è

$$\lim_{M \rightarrow \infty} E[S_{xx}(\omega)] = P_{xx}(\omega) \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\omega}) d\omega$$

Quindi, affinché la stima smussata dello spettro sia asintoticamente non polarizzata, dobbiamo imporre che sia

$$w(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(e^{j\omega}) d\omega = 1$$

In base all'espressione (11.44) la varianza del periodogramma risulta approssimativamente

$$\text{var}[I_N(\omega)] \cong \begin{cases} 2P_{xx}^2(0), & \omega = 0 \\ 2P_{xx}^2(\pi), & \omega = \pi \\ P_{xx}^2(\omega), & \text{altrove} \end{cases}$$

Perciò per $0 < \omega < \pi$, la varianza del periodogramma smussato $S_{xx}(\omega)$ differisce dalla varianza di $I_N(\omega)$ per il fattore

$$\frac{1}{N} \sum_{m=-(M-1)}^{M-1} w^2(m) = \frac{1}{2\pi N} \int_{-\pi}^{\pi} W^2(e^{j\omega}) d\omega \quad (11.60)$$

È chiaro quindi che nel progettare la stima bisogna scegliere M e la forma della finestra in modo che la varianza di $S_{xx}(\omega)$ sia minore della varianza di $I_N(\omega)$: in altri termini, il fattore (11.60) deve risultare minore di uno. Nel probl. 4 di questo capitolo verrà calcolato questo fattore per diverse finestre di uso comune.

11.4.3 Metodo di Welch - Media di periodogrammi modificati

Welch [9] ha introdotto un'utile modifica al metodo di Bartlett che, tra l'altro, si presta per il calcolo diretto di una stima dello spettro di potenza usando la FFT. La sequenza dati è ancora suddivisa in $K = N/M$ sottosequenze di M campioni ciascuna in base alla (11.45). In questo caso però la finestra $w(n)$ viene applicata direttamente sulle sottosequenze dei dati, prima del calcolo del periodogramma. Definiamo quindi i K periodogrammi modificati

$$J_M^{(i)}(\omega) = \frac{1}{MU} \left| \sum_{n=0}^{M-1} x^{(i)}(n)w(n)e^{-j\omega n} \right|^2, \quad i = 1, 2, \dots, K \quad (11.61)$$

dove è

$$U = \frac{1}{M} \sum_{n=0}^{M-1} w^2(n) \quad (11.62)$$

La stima dello spettro è definita come

$$B_{xx}^w(\omega) = \frac{1}{K} \sum_{i=1}^K J_M^{(i)}(\omega) \quad (11.63)$$

Si può dimostrare (v. probl. 5 di questo capitolo) che il valore atteso di $B_{xx}^w(\omega)$ è

$$E[B_{xx}^w(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\theta) W(e^{j(\omega-\theta)}) d\theta \quad (11.64)$$

con

$$W(e^{j\omega}) = \frac{1}{MU} \left| \sum_{n=0}^{M-1} w(n)e^{-j\omega n} \right|^2 \quad (11.65)$$

Il fattore di normalizzazione U è necessario affinché la stima $B_{xx}^w(\omega)$ sia asintoticamente non polarizzata. In [9] Welch dimostra che, se le sottosequenze di $x(n)$ sono disgiunte, risulta

$$\text{var}[B_{xx}^w(\omega)] \approx \frac{1}{K} P_{xx}^2(\omega)$$

come abbiamo già visto per il metodo di Bartlett. Sempre in [9] Welch esamina anche la possibilità di introdurre una sovrapposizione tra le sot-

tosequenze $x^{(i)}(n)$, per cui aumenta il numero dei periodogrammi modificati, che però non sono più indipendenti. Quindi applicando la finestra alle sottosequenze dei dati prima di calcolare il periodogramma, da un lato si ottiene la riduzione di varianza del metodo originale di Bartlett e dall'altro lo spettro viene smussato (cambia cioè la risoluzione della stima) in un modo che dipende dalla scelta della finestra dati. In questo caso la finestra spettrale è proporzionale al modulo quadrato della trasformata di Fourier della finestra e non semplicemente alla trasformata. Questo vuol dire che, qualunque sia la finestra applicata ai dati, la finestra spettrale corrispondente sarà sempre non negativa, e si può vedere che anche la stima dello spettro $B_{xx}^w(\omega)$ risulta sempre non negativa.

11.5 STIMA DELLA COVARIANZA INCROCIATA E DELLO SPETTRO INCROCIATO

Per stimare la funzione di covarianza incrociata e lo spettro incrociato di due diversi processi aleatori è possibile usare, con lievi modifiche, i metodi dei paragrafi precedenti. Siano, ad esempio, $x(n)$ e $y(n)$ due processi aleatori a media nulla, per cui $\gamma_{xy}(m) = \phi_{xy}(m)$. Allora la stima della covarianza (o correlazione) incrociata, corrispondente alla stima dell'autocovarianza $c_{xx}(m)$ espressa dalla (11.19), è

$$c_{xy}(m) = \frac{1}{N} \sum_{n=0}^{N-m-1} x(n)y(n+m), \quad 0 \leq m < N \quad (11.66a)$$

$$c_{xy}(-m) = \frac{1}{N} \sum_{n=0}^{N-m-1} x(n+m)y(n), \quad 0 \leq m < N \quad (11.66b)$$

Si noti che quando è $y(n) = x(n)$, la formula (11.66) si riduce alla (11.19).

Il valore atteso dell'espressione (11.66a) è

$$E[c_{xy}(m)] = \left(1 - \frac{m}{N}\right) \phi_{xy}(m), \quad 0 \leq m < N$$

dove $\phi_{xy}(m)$ è la sequenza di correlazione incrociata vera. Analogamente si ha per la (11.66b)

$$\begin{aligned} E[c_{xy}(-m)] &= \left(1 - \frac{m}{N}\right) \phi_{yx}(m) \\ &= \left(1 - \frac{m}{N}\right) \phi_{xy}(-m) \quad 0 \leq m < N \end{aligned}$$

Combinando queste due relazioni si ottiene

$$E[c_{xy}(m)] = \left(1 - \frac{|m|}{N}\right) \phi_{xy}(m), \quad -N < m < N \quad (11.67)$$

Si vede che $c_{xy}(m)$ è una stima asintoticamente non polarizzata della covarianza incrociata $\phi_{xy}(m)$. Come per la stima $c_{xx}(m)$, la varianza della stima $c_{xy}(m)$ è inversamente proporzionale a N .

Per stimare lo spettro di potenza incrociato possiamo eseguire la trasformata di Fourier di $c_{xy}(m)$ e ottenere la stima

$$C_{xy}(\omega) = \sum_{m=-(N-1)}^{N-1} c_{xy}(m) e^{-j\omega m} \quad (11.68)$$

Se è $x(n) = y(n)$, si vede dalla (11.24) che l'espressione (11.68) si riduce al periodogramma. Si noti che in generale $c_{xy}(m)$ non ha proprietà di simmetria particolari, per cui $C_{xy}(\omega)$ è di solito una funzione complessa. Si dimostra facilmente che è

$$E[C_{xy}(\omega)] = \sum_{m=-(N-1)}^{N-1} \left(1 - \frac{|m|}{N}\right) \phi_{xy}(m) e^{-j\omega m} \quad (11.69)$$

Anche se si vede dalla formula precedente che $C_{xy}(\omega)$ è una stima asintoticamente non polarizzata dello spettro di potenza incrociato, proprio come nel caso del periodogramma, la varianza di $C_{xy}(\omega)$ non tende a zero al crescere di N . Per ridurre la varianza e smussare la stima occorre quindi applicare delle finestre o fare la media su stime ottenute da sottosequenze. Per esempio, consideriamo la stima smussata dello spettro

$$S_{xy}(\omega) = \sum_{m=-(M-1)}^{M-1} c_{xy}(m) w(m) e^{-j\omega m} \quad (11.70)$$

Sotto ipotesi simili a quelle usate nel par. 11.4 si può dimostrare che risulta

$$E[S_{xy}(\omega)] \cong \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xy}(\theta) W(e^{j(\omega-\theta)}) d\theta \quad (11.71)$$

Analogamente, si può dimostrare che $\text{var}[S_{xy}(\omega)]$ diminuisce al crescere della lunghezza della sequenza e anche al diminuire della lunghezza della finestra $w(m)$ della (11.70). Perciò il problema di conciliare la risoluzione dello spettro e la riduzione di varianza è del tutto simile a quello già discusso per lo spettro di potenza di un solo segnale aleatorio.

Noi non ci addentreremo ulteriormente nei dettagli riguardanti le stime della covarianza e dello spettro incrociati. In [5] si trovano diversi capitoli dedicati a queste stime e alle loro applicazioni.

11.6 USO DELLA FFT NELLA STIMA DELLO SPETTRO

La FFT costituisce un mezzo efficiente per calcolare stime dello spettro di potenza a frequenze equispaziate $\omega_k = (2\pi/M)k$. Inoltre, usando le tecniche sviluppate nel cap. 3 per il calcolo della convoluzione, si può usare la FFT per ottenere stime della covarianza. In questo paragrafo esamineremo i dettagli di questi metodi di calcolo.

11.6.1 Applicazione ai metodi di Bartlett o di Welch

Supponiamo di voler calcolare una stima dello spettro a frequenze equispaziate facendo la media di periodogrammi, come illustrato nel paragrafo 11.4.3. Desideriamo cioè calcolare

$$B_{xx}^w\left(\frac{2\pi}{M}k\right) = \frac{1}{K} \sum_{i=1}^K J_M^{(i)}\left(\frac{2\pi}{M}k\right), \quad k = 0, 1, \dots, M-1$$

dove è

$$J_M^{(i)}\left(\frac{2\pi}{M}k\right) = \frac{1}{MU} \left| \sum_{n=0}^{M-1} x^{(i)}(n) w(n) e^{-j(2\pi/M)kn} \right|^2$$

per $i = 1, 2, \dots, K$ e per $k = 0, 1, \dots, M-1$. Lo si può fare usando la seguente procedura. Per ogni sottosequenza si calcola

$$X_M^{(i)}(k) = \sum_{n=0}^{M-1} x^{(i)}(n) w(n) e^{-j(2\pi/M)kn}, \quad k = 0, 1, \dots, M-1$$

mediante un opportuno algoritmo di FFT² e si ricavano le quantità $|X_M^{(i)}(k)|^2$. Queste ultime si sommano una all'altra fino a $i = K$ e il risultato finale si divide per KMU . Se i dati $x(n)$ sono reali si possono ricavare due trasformate in una sola volta, sfruttando le proprietà di simmetria illustrate nel probl. 10 del cap. 6, e riducendo quindi notevolmente il peso dei calcoli.

Questo è un procedimento molto semplice e il risultato è una stima diretta dello spettro di potenza che sarà sempre non negativa e che possiamo interpretare come abbiamo fatto nei par. 11.4.1 e 11.4.3. Se però vogliamo stimare anche la funzione di correlazione insieme allo spettro di potenza, è meglio calcolare prima la stima $c_{xx}(m)$ della correlazione e poi la stima dello spettro, poiché la semplice operazione di antitrasformare $B_{xx}^w((2\pi/M)k)$ con la DFT porta a qualcosa che possiamo chiamare, nel caso migliore, una stima con *aliasing* temporale della sequenza di correlazione (v. probl. 6 di questo capitolo). Perciò esamineremo ora come si può usare la FFT per calcolare stime di quest'ultima.

11.6.2 Calcolo di stime della correlazione

La FFT può essere usata per calcolare in maniera efficiente la stima dell'autocorrelazione

$$c_{xx}(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n)x(n+m), \quad 0 \leq m \leq M-1 \quad (11.72)$$

dove è $M \leq N$ [se da $x(n)$ si sottrae prima la media campione, $c_{xx}(m)$ è una stima dell'autocovarianza]. Ricordando che è $c_{xx}(-m) = c_{xx}(m)$, è chiaro che basta valutare l'espressione precedente per m positivi. Il punto

² Si noti che la finestra $w(n) = 1$ per $0 \leq n \leq M-1$ corrisponde al metodo di Bartlett del par. 11.4.1, mentre qualsiasi altra scelta di $w(n)$ corrisponde al metodo di Welch del par. 11.4.3.

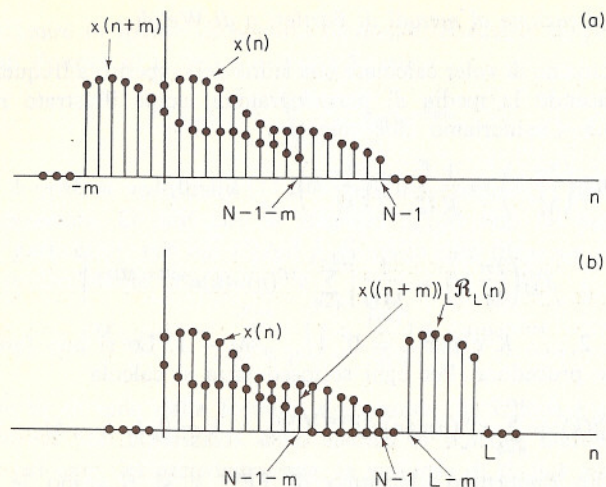


Fig. 11.4 Calcolo della stima dell'autocorrelazione: (a) $x(n)$ e $x(n+m)$ per una sequenza lunga N ; (b) sequenze periodiche $x(n)$ e $x(n+m)$ di cui si fa la correlazione circolare.

chiave per capire come si possono usare le trasformate di Fourier veloci per calcolare $c_{xx}(m)$ sta nell'osservazione che $c_{xx}(m)$ è la convoluzione discreta di $x(n)$ con $x(-n)$. Supponiamo di calcolare $X(k)$, la DFT di $x(n)$, e di moltiplicarla per $X^*(k)$. La DFT inversa di $X(k)X^*(k) = |X(k)|^2$ corrisponde alla convoluzione circolare di $x(n)$ con $x(-n)$, cioè a una correlazione circolare. Possiamo ottenere che i valori della correlazione circolare siano corretti (cioè uguali a quelli della correlazione « lineare ») nell'intervallo $0 \leq m \leq M-1$, allungando la sequenza $x(n)$ con $(L-N)$ zeri e calcolando una DFT su L punti.

Per chiarire come va scelto L , consideriamo la fig. 11.4. In fig. 11.4(a) sono rappresentate le due sequenze $x(n)$ e $x(n+m)$ per un particolare valore (positivo) di m . La fig. 11.4(b) mostra le sequenze $x(n)$ e $x((n+m))_L$ di cui si fa la correlazione circolare corrispondente a $|X(k)|^2$. È chiaro che la correlazione circolare sarà uguale a $Nc_{xx}(m)$ per $0 \leq m \leq M-1$ se $x((n+m))_L$ ripetendosi non si sovrappone a $x(n)$ per $0 \leq m \leq M-1$. Si vede dalla fig. 11.4(b) che questo si verifica se è $L \geq N + M - 1$.

Quindi il procedimento per calcolare $c_{xx}(m)$ per $0 \leq m \leq M-1$ è il seguente:

1. Si costruisce una sequenza di L punti aggiungendo a $x(n)$ $(M-1)$ zeri.
2. Si calcola la DFT su L punti

$$X(k) = \sum_{n=0}^{L-1} x(n)e^{-j(2\pi/L)kn}, \quad k = 0, 1, \dots, L-1$$

3. Si calcola la DFT inversa su L punti

$$v(m) = \frac{1}{L} \sum_{k=0}^{L-1} |X(k)|^2 e^{j(2\pi/N)km}, \quad m = 0, 1, \dots, L-1$$

4. Infine

$$c_{xx}(m) = \frac{1}{N} v(m), \quad m = 0, 1, \dots, M-1$$

Se M è piccolo, il metodo più efficiente può essere l'uso diretto della relazione (11.72): la quantità di calcoli è allora proporzionale a $N \cdot M$. Invece, il procedimento descritto prima richiede un numero di operazioni proporzionale a

$$L \log L = (N + M) \log (N + M)$$

per cui risulta più efficiente per valori di M abbastanza grandi. Il valore esatto di M che separa i due campi dipende naturalmente dalla particolare struttura usata per il calcolo della DFT; è lecito comunque attendersi che questo valore sia molto inferiore a 100 [10].

Abbiamo visto che per ridurre la varianza della stima $c_{xx}(m)$ occorre rendere N grande. Può risultare allora difficile, se non impossibile, calcolare in modo efficiente la DFT su L punti che è richiesta in questo caso. Tuttavia, essendo di solito M molto minore di N , possiamo suddividere l'ingresso in sottosequenze, secondo un procedimento simile a quelli discussi nel par. 3.9 per la convoluzione.

Per chiarire il metodo, riscriviamo l'espressione (11.72) come

$$c_{xx}(m) = \frac{1}{N} \left[\sum_{n=0}^{M-1} x(n)x(n+m) + \sum_{n=M}^{2M-1} x(n)x(n+m) + \dots + \sum_{n=(K-1)M}^{KM-1} x(n)x(n+m) \right]$$

dove è $N = KM$. Questa trasformazione è possibile per il fatto che il limite superiore che compare nella (11.72) può essere sostituito da $N-1$, se si considera nulla $x(n)$ al di fuori dell'intervallo $0 \leq n \leq N-1$. Facendo le opportune sostituzioni possiamo scrivere

$$c_{xx}(m) = \frac{1}{N} \sum_{i=1}^K \sum_{n=0}^{M-1} x(n + (i-1)M)x(n + (i-1)M + m)$$

e definendo

$$v_i(m) = \sum_{n=0}^{M-1} x(n + (i-1)M)x(n + (i-1)M + m) \quad (11.73)$$

risulta

$$c_{xx}(m) = \frac{1}{N} \sum_{i=1}^K v_i(m), \quad 0 \leq m \leq M-1 \quad (11.74)$$

Per valutare l'espressione (11.73) è opportuno definire le sequenze lunghe L

$$x_i(n) = \begin{cases} x(n + (i-1)M), & 0 \leq n \leq M-1 \\ 0, & M \leq n \leq L-1 \end{cases} \quad (11.75a)$$

e

$$y_i(n) = x(n + (i-1)M), \quad 0 \leq n \leq L-1 \quad (11.75b)$$

Allora la correlazione circolare

$$v_i(m) = \sum_{n=0}^{M-1} x_i(n) y_i((n+m))_L$$

è uguale a $v(m)$ per $0 \leq m \leq M-1$ se è $L \geq 2M-1$. Nella fig. 11.5 sono riportati due tipici esempi di queste sequenze. Se $X_i(k)$ e $Y_i(k)$ sono le DFT su L punti di $x_i(n)$ e $y_i(n)$, allora si ha, per $L \geq 2M-1$,

$$v_i(m) = \frac{1}{L} \sum_{k=0}^{L-1} V_i(k) W_L^{km} \quad 0 \leq m \leq M-1$$

dove è

$$V_i(k) = X_i(k) Y_i^*(k) \quad (11.76)$$

Se invece dell'espressione (11.74) calcoliamo

$$V(k) = \sum_{i=1}^K V_i(k), \quad k = 0, 1, \dots, L-1 \quad (11.77)$$

allora si ha

$$c_{xx}(m) = \frac{1}{N} v(m) = \frac{1}{L} \sum_{k=0}^{L-1} V(k) e^{j(2\pi/L)km}, \quad 0 \leq m \leq M-1 \quad (11.78)$$

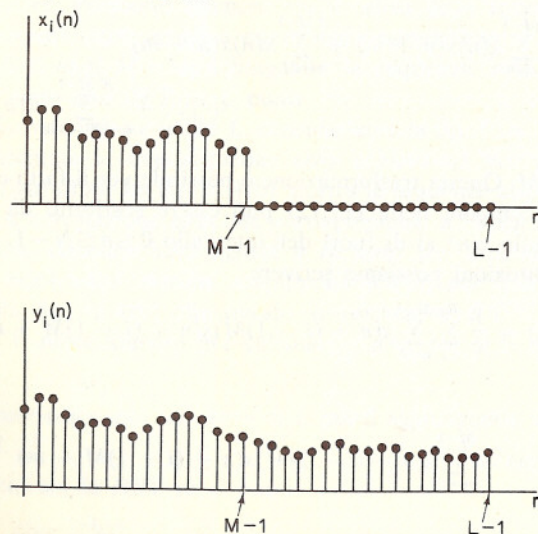


Fig. 11.5 Le due sequenze di L punti necessarie per calcolare il contributo del segmento i -esimo alla stima dell'autocorrelazione.

Quindi possiamo ottenere M valori di $c_{xx}(m)$ calcolando $2K$ DFT lunghe L ed una DFT inversa lunga ancora L .

Si noti che il procedimento descritto vale anche per il calcolo della stima della correlazione incrociata $c_{xy}(m)$. Supponiamo di avere due registrazioni lunghe N delle sequenze $x(n)$ e $y(n)$. Al posto delle (11.75) definiamo

$$x_i(n) = \begin{cases} x(n + (i-1)M), & 0 \leq n \leq M-1 \\ 0, & M \leq n \leq L-1 \end{cases} \quad (11.79a)$$

come prima, e invece

$$y_i(n) = y(n + (i-1)M), \quad 0 \leq n \leq L-1 \quad (11.79b)$$

per $i = 1, 2, \dots, K$. Se allora usiamo, come prima, le relazioni (11.76) e (11.77), otteniamo

$$c_{xy}(m) = \frac{1}{N} v(m), \quad 0 \leq m \leq M-1 \quad (11.80)$$

Allo stesso modo si calcola $c_{xy}(m)$ per $m < 0$, scambiando solo le sequenze x e y , essendo $c_{yx}(m) = c_{xy}(-m)$.

Rader [10] ha dimostrato che con la scelta particolare $L = 2M$ è possibile risparmiare notevolmente nel calcolo delle stime dell'autocorrelazione. La fig. 11.6 presenta due insiemi di sequenze $x_i(n)$, $y_i(n)$ e $x_{i+1}(n)$, $y_{i+1}(n)$ con $L=2M$ come richiesto per il calcolo dell'autocorrelazione. Dalla figura risulta chiaro che è

$$y_i(n) = x_i(n) + x_{i+1}(n - M) \quad (11.81)$$

Da questa relazione segue che è

$$Y_i(k) = X_i(k) + (-1)^k X_{i+1}(k), \quad k = 0, 1, \dots, 2M-1 \quad (11.82)$$

Perciò $Y_i(k)$ può essere ottenuta usando la (11.82) invece che calcolando a parte una FFT. Inoltre, con una sola FFT si possono ricavare due trasformate, ad es. $X_i(k)$ e $X_{i+1}(k)$, usando le tecniche discusse nel probl. 10 del cap. 6. Quindi, il procedimento per il calcolo di $c_{xx}(m)$ può essere così riassunto:

1. Si costruisce la sequenza

$$x_1(n) = \begin{cases} x(n), & 0 \leq n \leq M-1 \\ 0, & M \leq n \leq 2M-1 \end{cases}$$

e si calcola la trasformata $X_1(k)$ su $2M$ punti. Poniamo $A_0(k) = 0$.

2. Per $i = 1, 2, \dots, K$, si pone

$$x_{i+1}(n) = \begin{cases} x(n + iM), & 0 \leq n \leq M-1 \\ 0, & M \leq n \leq 2M-1 \end{cases}$$

e si calcola la trasformata $X_{i+1}(k)$ su $2M$ punti. Definiamo $X_{K+1}(k) = 0$.

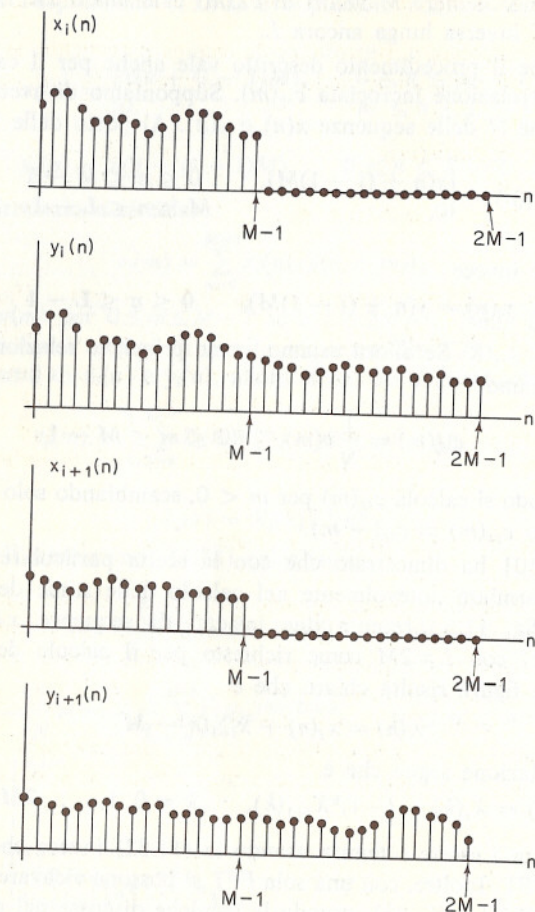


Fig. 11.6 Illustrazione del fatto che la prima metà di $x_{i+1}(n)$ è identica alla seconda metà di $y_i(n)$ quando è $L = 2M$.

Si calcola poi, per $0 \leq k \leq 2M-1$ e per $i=1, 2, \dots, K$,

$$A_i(k) = A_{i-1}(k) + X_i(k)[X_i^*(k) + (-1)^k X_{i+1}^*(k)]$$

3. Infine, definendo $V(k) = A_K(k)$ e indicando con $v(m)$ la DFT inversa di $V(k)$ lunga $2M$, si ha

$$c_{xx}(m) = \frac{1}{N} v(m), \quad 0 \leq m \leq M$$

Quindi, nel caso particolare di $L = 2M$ e solo per il calcolo di $c_{xx}(m)$, occorre calcolare K trasformate $X_i(k)$ ed una trasformata inversa.

11.6.3 Calcolo di stime di spettro smussate a partire da $c_{xx}(m)$

Una volta calcolata $c_{xx}(m)$ usando la tecnica precedente, si possono ottenere i campioni della stima dello spettro smussata $S_{xx}(\omega)$ costruendo la sequenza

$$s_{xx}(m) = \begin{cases} c_{xx}(m)w(m), & 0 \leq m \leq M-1 \\ 0, & M \leq m \leq L-M \\ c_{xx}(L-m)w(L-m), & L-M+1 \leq m \leq L-1 \end{cases} \quad (11.83)$$

dove $w(m)$ è una finestra opportuna. La DFT di $s_{xx}(m)$ è allora

$$S_{xx}(k) = S_{xx}(\omega)|_{\omega=(2\pi/L)k}, \quad k = 0, 1, \dots, L-1$$

Si noti che L può essere scelto arbitrariamente grande (l'unico vincolo è quello connesso all'esecuzione dei calcoli), ottenendo così campioni di $S_{xx}(\omega)$ molto ravvicinati in frequenza; l'effettiva risoluzione frequenziale rimane però determinata dalla forma e dalla lunghezza della finestra $w(m)$.

11.7 ESEMPIO DI STIMA DELLO SPETTRO

Nel cap. 9 abbiamo fatto l'ipotesi che l'errore dovuto alla quantizzazione sia un rumore bianco e, inoltre, che questo rumore di quantizzazione sia scorrelato rispetto al segnale originale. Possiamo adesso verificare la validità di queste ipotesi stimando le sequenze di covarianza e gli spettri di potenza con i metodi illustrati in questo capitolo.

Prendiamo come esempio l'esperimento rappresentato in fig. 11.7. Un segnale vocale $x_a(t)$ già filtrato con un passa-basso è stato campionato

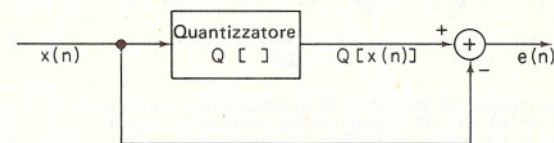


Fig. 11.7 Esperimento per determinare le proprietà del rumore di quantizzazione.

alla frequenza di 10 KHz, fornendo la sequenza di campioni $x(n)$ (assumeremo che i campioni risultanti abbiano precisione infinita). Il campo di variazione del valore dei campioni è

$$-16,000 \leq x(n) \leq 16,000$$

Questi campioni sono quantizzati con un quantizzatore lineare a otto bit, dopo di che viene calcolata la sequenza errore

$$e(n) = Q[x(n)] - x(n)$$

La fig. 11.8(a) mostra 400 campioni consecutivi del segnale vocale e la fig. 11.8(b) riporta la sequenza errore corrispondente (i campioni sono

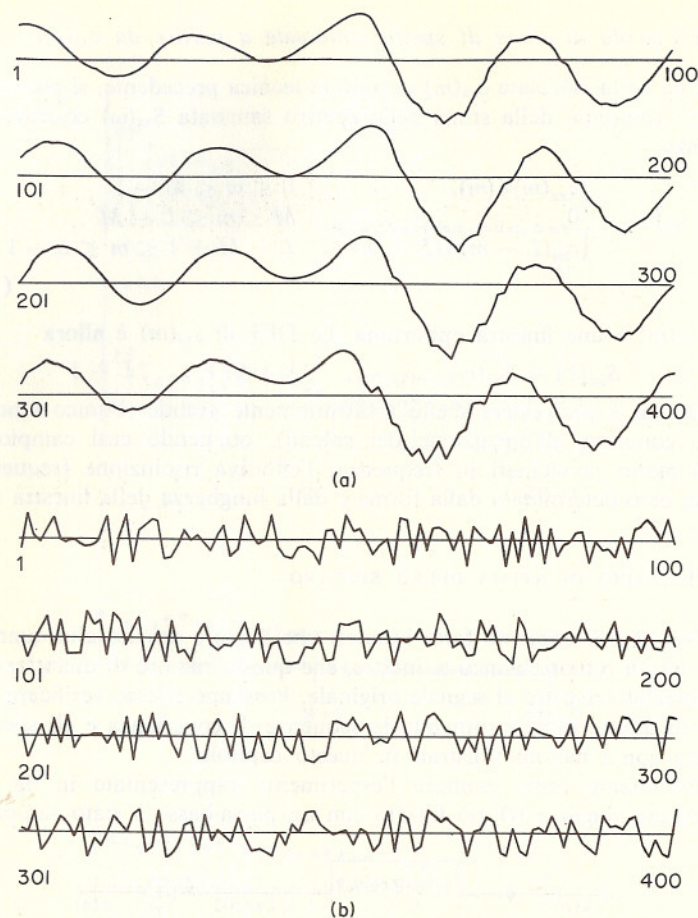


Fig. 11.8 Segnale vocale (a) ed errore di quantizzazione corrispondente (b) per una quantizzazione ad otto bit (ingrandimento di 66 volte rispetto ad (a)). Ogni riga corrisponde a 100 campioni consecutivi congiunti da tratti di retta per comodità di rappresentazione.

congiunti con tratti di retta per comodità di rappresentazione). La prima impressione che si ricava dal confronto di questi due grafici tende a confermare la validità delle ipotesi enunciate prima, anche se da un esame più dettagliato può apparire un certo grado di correlazione.

La fig. 11.9 mostra una stima della autocovarianza normalizzata e dello spettro di potenza della sequenza errore, relativa ad una lunghezza di 2000 campioni. La media e la varianza di $e(n)$ sono state stimate con i metodi discussi nel par. 11.1 e queste stime sono risultate $\hat{m}_e = -1$ e $\hat{\sigma}_e^2 = 1300$. Si è poi calcolata la stima della covarianza per $M = 512$ usando il metodo del par. 11.6.2, dopo aver sottratto \hat{m}_e da $e(n)$. Infine,

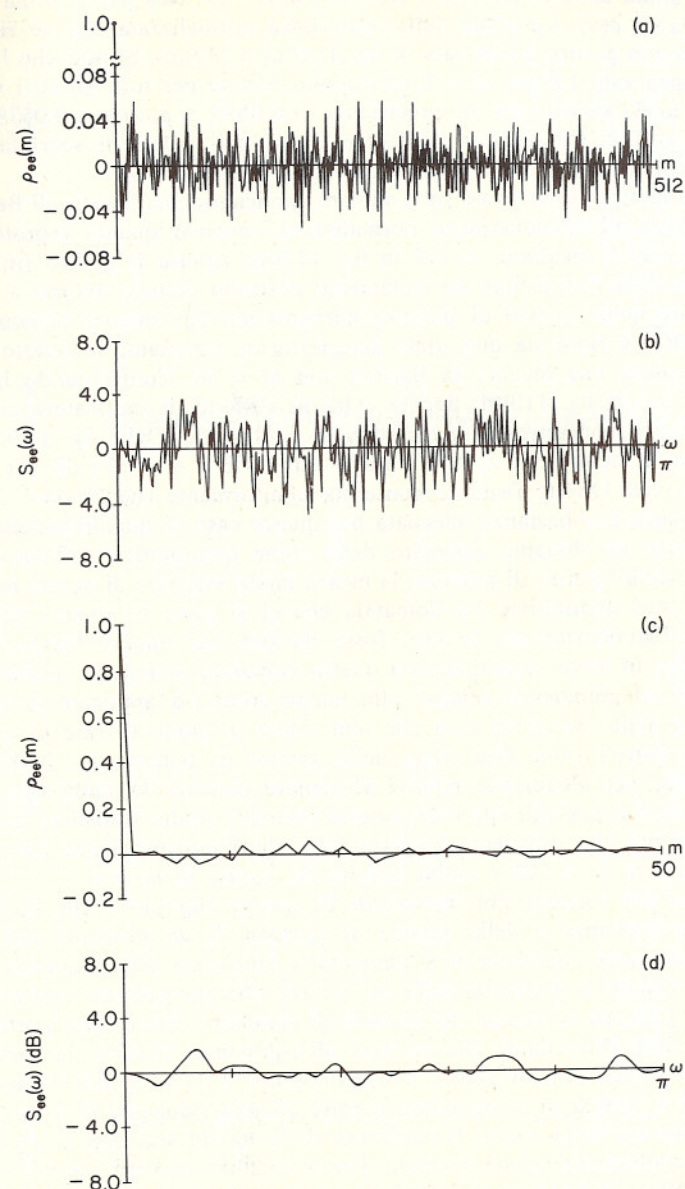


Fig. 11.9 (a) stima dell'autocovarianza normalizzata del rumore introdotto da una quantizzazione a otto bit; la sequenza analizzata è lunga 2000 punti; (b) stima dello spettro di potenza usando la finestra di Bartlett, con $M = 512$; (c) stima dell'autocovarianza normalizzata, $0 \leq M \leq 50$; (d) stima dello spettro di potenza usando la finestra di Bartlett, con $M = 50$.

questa stima della covarianza è stata divisa per σ_e^2 , cioè per la stima della varianza di $e(n)$. La stima della covarianza normalizzata che ne risulta, indicata con $\rho_{ee}(m)$, è mostrata in fig. 11.9(a) e 11.9(c). Si noti che l'autocovarianza vale 1.0 per $m = 0$ ed è molto minore per tutti gli altri valori di m ; anzi, risulta per la precisione $-0.0548 \leq \rho_{ee}(m) \leq 0.0579$ per $1 \leq m \leq 512$. Ciò sembra confortare la nostra ipotesi di scorrelazione tra campione e campione della sequenza errore.

Lo spettro di potenza è stato stimato applicando una finestra di Bartlett ($M = 512$) all'autocovarianza normalizzata, secondo quanto esposto nel par. 11.6.3. Il risultato, di cui la fig. 11.9(b) riporta il grafico (in dB), presenta delle fluttuazioni ad andamento piuttosto casuale intorno a 0 dB (il valore dello spettro di potenza normalizzato del rumore bianco). In fig. 11.9(d) è riportata una stima maggiormente smussata: in questo caso è stata usata una finestra di Bartlett con $M = 50$. Confrontando la fig. 11.9(b) con la fig. 11.9(d) appare evidente l'effetto di smussatura che ne risulta, corrispondente ad una perdita di risoluzione. Dalla fig. 11.9(d) si vede che la stima dello spettro è tra -1.097 dB e $+1.631$ dB per tutte le frequenze. Questo risultato conferma ulteriormente che l'ipotesi di rumore bianco è abbastanza adeguata per questo caso di quantizzazione.

Anche se abbiamo calcolato delle stime quantitative dell'autocovarianza e dello spettro di potenza, la nostra interpretazione di queste misure è stata solo qualitativa. La domanda che ci si pone è: quanto piccola sarebbe l'autocovarianza se $e(n)$ fosse davvero un rumore bianco? Per rispondere in modo quantitativo a questa domanda potremmo ricavare gli intervalli di confidenza relativi alle nostre stime ed applicare la teoria statistica della decisione, cosa che non siamo in grado di fare in questa sintetica introduzione alla stima dello spettro di potenza (v. Jenkins e Watts [5] per alcuni test relativi al rumore bianco). In molti casi, tuttavia, non è necessario effettuare questo ulteriore studio statistico: spesso, infatti, è più che sufficiente l'osservazione che l'autocovarianza normalizzata per $1 \leq m \leq 512$ è molto minore del valore in $m = 0$.

Uno dei concetti più importanti di questo capitolo è che la stima dell'autocovarianza e dello spettro di potenza di un processo aleatorio stazionario deve migliorare se si aumenta la lunghezza della sequenza analizzata. Questo è illustrato dalla fig. 11.10, che corrisponde esattamente alla fig. 11.9, ad eccezione del numero di campioni, che è stato portato a $N = 14.000$. Ricordiamo che, in base all'espressione (11.23), la varianza della stima dell'autocovarianza è proporzionale a $1/N$: perciò l'aumento da 2000 a 14.000 di N dovrebbe portare ad una riduzione di sette volte della varianza della stima. Il confronto della fig. 11.9(a) con la 11.10(a) sembra confermare questo risultato. Per $N = 2000$ la stima assume valori tra i limiti -0.0548 e $+0.0579$, mentre per $N = 14.000$ i limiti sono -0.0254 e $+0.0239$. Questo concorda, intuitivamente, con la diminuzione che ci aspettavamo di un fattore sette nella varianza. Notiamo che in base all'espressione (11.58) ci si deve attendere una riduzione simile per la varianza della stima dello spettro. Anche questo fatto risulta evidente

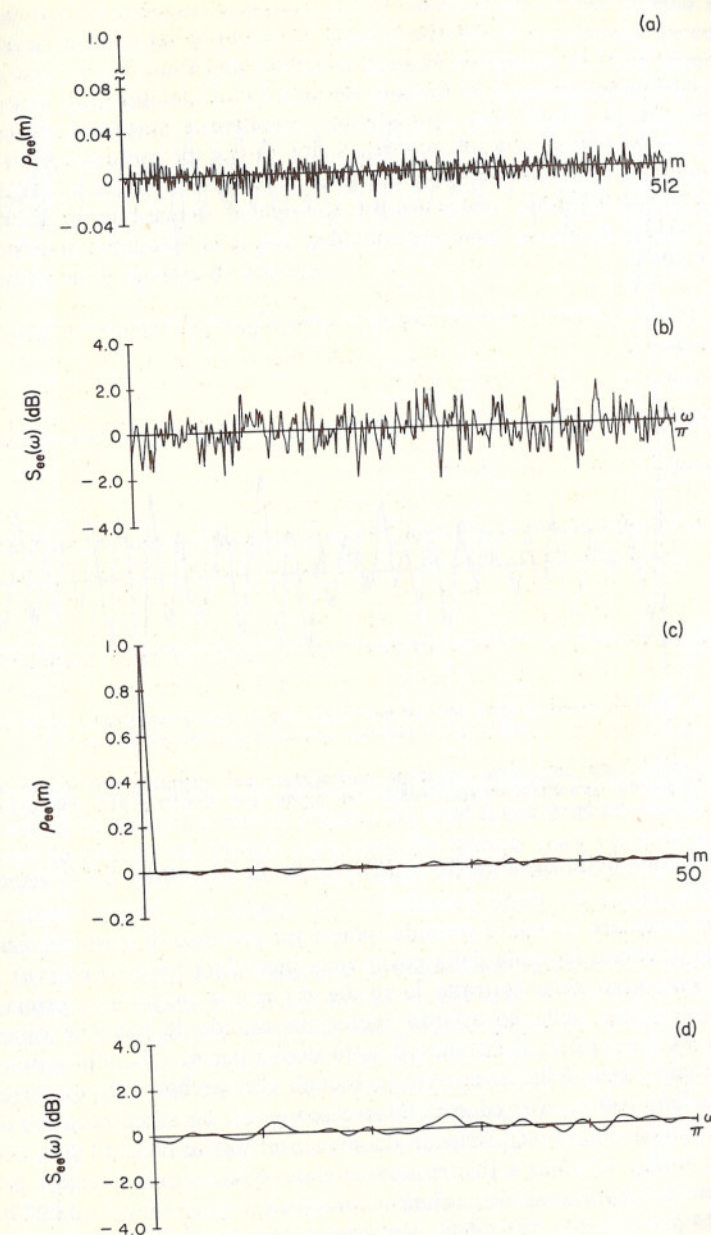


Fig. 11.10 (a) stima dell'autocovarianza normalizzata del rumore introdotto da una quantizzazione a otto bit; la sequenza analizzata è lunga 14.000 punti; (b) stima dello spettro di potenza usando la finestra di Bartlett, con $M=512$; (c) stima dell'autocovarianza normalizzata, $0 \leq m \leq 50$; (d) stima dello spettro di potenza usando la finestra di Bartlett, con $M=50$.

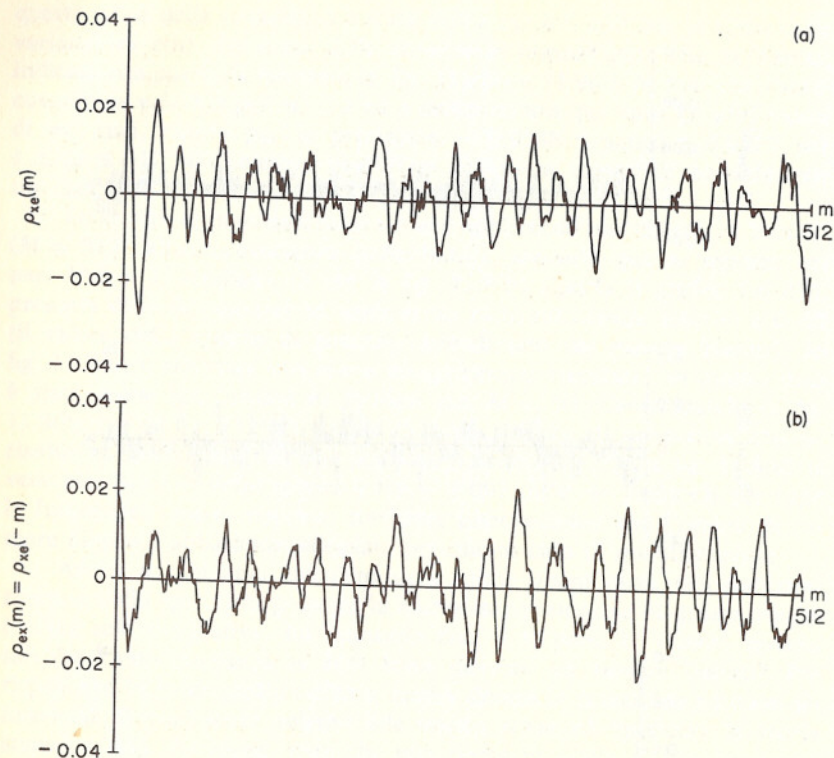


Fig. 11.11 Stima della covarianza incrociata normalizzata per quantizzazione a otto bit e sequenza analizzata lunga 14.000: (a) $\rho_{xe}(m)$ per $0 \leq m \leq 511$; (b) $\rho_{ex}(m) = \rho_{xe}(-m)$ per $0 \leq m \leq 511$.

dal confronto delle fig. 11.9(b) e (d) con le fig. 11.10(b) e (d) rispettivamente.

Per verificare la nostra seconda ipotesi sul processo di quantizzazione, abbiamo calcolato la stima della covarianza incrociata tra $x(n)$ ed $e(n)$. In questo caso sono state sottratte le medie $\hat{m}_x = 1$ e $\hat{m}_e = -1$, prima di eseguire le stime della covarianza incrociata usando le tecniche espone nei par. 11.5 e 11.6.2. Il risultato è stato diviso per $\hat{\sigma}_x \cdot \hat{\sigma}_e$, allo scopo di ottenere una stima della covarianza incrociata che sarebbe 1.0 per correlazione perfetta e 0 se $x(n)$ ed $e(n)$ fossero scorrelati. Le stime normalizzate $\rho_{xe}(m)$ e $\rho_{ex}(m) = \rho_{xe}(-m)$, ricavate da sequenze lunghe $N = 14.000$, sono mosirate in fig. 11.11(a) e (b), rispettivamente. Notiamo che i valori della sequenza di covarianza incrociata normalizzata sono tra -0.0279 e $+0.0222$ per il $-511 \leq m \leq 511$. Ricordando che la covarianza incrociata normalizzata vale uno quando la correlazione è perfetta, siamo spinti ancora una volta ad accettare l'ipotesi che il rumore di quantizzazione sia scorrelato dall'ingresso del quantizzatore.

Nel cap. 9 avevamo concluso che il modello di rumore bianco era accettabile fintantoché il passo di quantizzazione si manteneva piccolo. Questa condizione non è più verificata quando il numero di bit è piccolo. Per vedere come cambia lo spettro del rumore di quantizzazione, è stato ripetuto l'esperimento precedente, usando solo otto livelli di quantizzazione, ovvero tre bit. La fig. 11.12 mostra l'errore di quantizzazione a tre bit relativo al segmento di segnale vocale di fig. 11.8(a). Si noti che in alcuni tratti l'andamento dell'errore appare molto simile a quello della forma d'onda originale: ci si può attendere che resti traccia di questo fatto nella stima dello spettro di potenza.

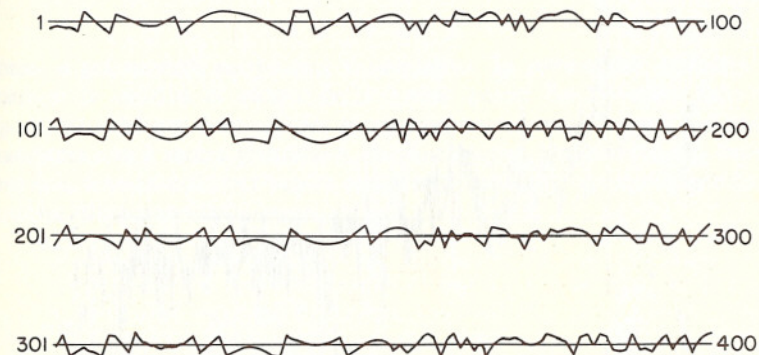


Fig. 11.12 Andamento dell'errore di quantizzazione per una quantizzazione a tre bit. La scala è la stessa del segnale originale, riportato in fig. 11.8(a).

La fig. 11.13 mostra le stime dell'autocovarianza e dello spettro di potenza della sequenza errore relativa ad una quantizzazione a tre bit e ad una lunghezza di 14.000 campioni. In questo caso l'autocovarianza riportata in fig. 11.13(a) e (c) è molto meno simile all'autocovarianza ideale del rumore bianco. La tabella 11.1 fornisce i primi 10 valori di $\rho_{xx}(m)$ per questo caso.

Le fig. 11.13(b) e (d) mostrano le stime dello spettro di potenza ottenute usando la finestra di Bartlett con $M = 512$ e $M = 50$, rispettivamente. È chiaro che lo spettro, in questo caso, non è piatto (in effetti, tende ad assumere la forma generale dello spettro della voce). Quindi, in questo caso, il modello del rumore di quantizzazione come rumore bianco può solo essere considerato un'approssimazione piuttosto grossolana.

La fig. 11.14 mostra la covarianza incrociata normalizzata tra il segnale $x(n)$ e il rumore di quantizzazione a tre bit $e(n)$. In questo caso la correlazione risulta un po' maggiore; tuttavia, possiamo ancora essere inclini ad accettare l'ipotesi che il segnale e l'errore di quantizzazione siano scorrelati.

Questo esempio illustra come vengano spesso usate le stime della covarianza e dello spettro di potenza per corroborare la nostra fiducia nelle

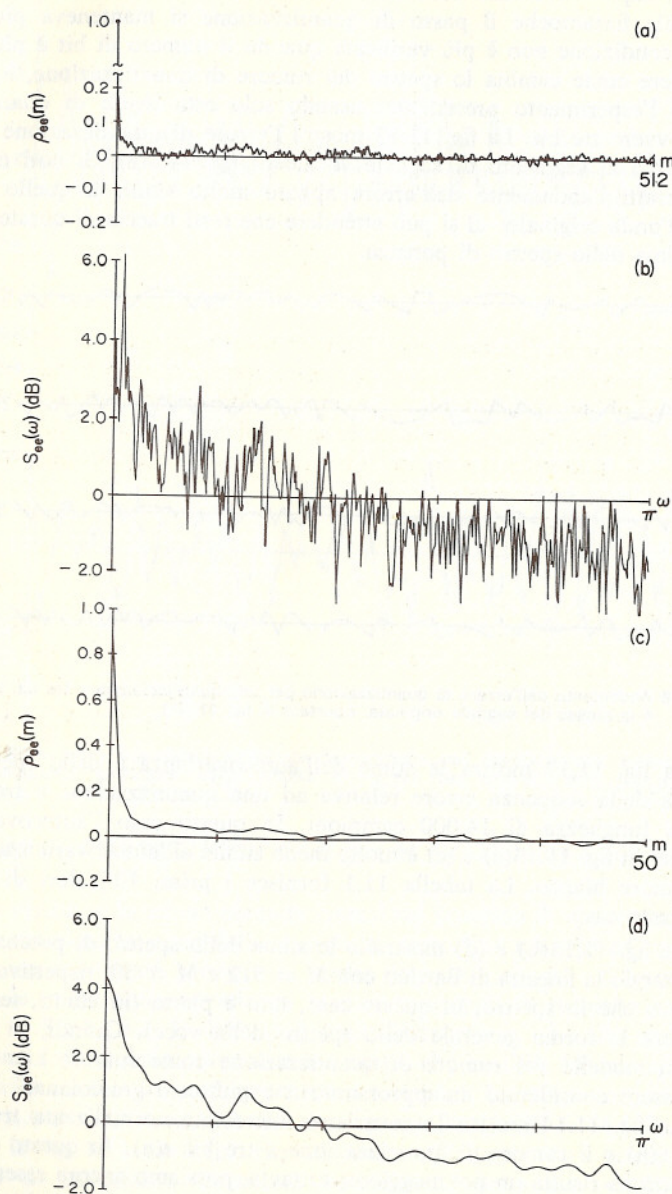


Fig. 11.13 (a) Stima dell'autocovarianza normalizzata per una quantizzazione a tre bit; la sequenza analizzata è lunga 14.000 punti; (b) stima dello spettro di potenza usando la finestra di Bartlett, con $M=512$; (c) stima dell'autocovarianza normalizzata, $0 \leq m \leq 50$; (d) stima dello spettro di potenza usando la finestra di Bartlett, con $M=50$.

Tab. 11.1

m	$\rho_{xx}(m)$
0	1.0
1	0.192
2	0.084
3	0.055
4	0.040
5	0.056
6	0.050
7	0.045
8	0.037
9	0.038

ipotesi o nei modelli teorici che si utilizzano. In particolare, abbiamo dimostrato la validità di alcune delle nostre ipotesi fondamentali fatte nel cap. 9, ed abbiamo dato un'idea di come queste ipotesi cadano quando la quantizzazione è molto grossolana. In conclusione, l'esempio fatto, benché semplice, è stato utile per capire come sono applicate spesso in pratica le tecniche di questo capitolo.

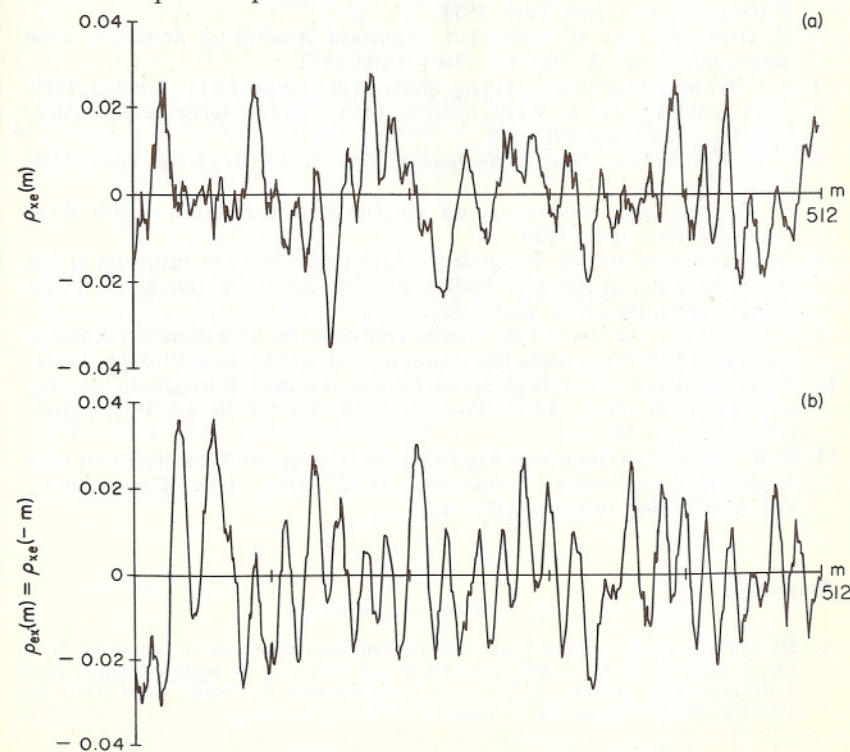


Fig. 11.14 Stima della covarianza incrociata normalizzata per una quantizzazione a tre bit: (a) $\rho_{xy}(m)$ per $0 \leq m \leq 511$; (b) $\rho_{yx}(m) = \rho_{xy}(-m)$ per $0 \leq m \leq 511$.

SOMMARIO

In questo capitolo abbiamo discusso alcune tecniche per la stima di medie di processi aleatori a tempo discreto. Il nostro scopo è stato quello di giustificare l'uso di alcuni concetti della teoria statistica della stima e di illustrarne le conseguenze. Ci siamo occupati in particolare dei risultati approssimati relativi alla media e alla varianza delle stime dell'autocovarianza e dello spettro di potenza di un processo aleatorio. Un argomento fondamentale è stato quello delle procedure di calcolo di queste stime e l'uso della FFT per renderle più efficienti. Un paragrafo finale ha illustrato l'applicazione di alcune delle tecniche descritte nel capitolo allo studio delle proprietà del rumore di quantizzazione.

BIBLIOGRAFIA

1. M. S. Bartlett, *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*, Cambridge University Press, New York, 1953.
2. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, Dover Publications, Inc., New York, 1958.
3. U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, John Wiley & Sons, Inc., New York, 1957.
4. E. J. Hannan, *Time Series Analysis*, Methuen & Company Ltd., London, 1960.
5. G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, Inc., San Francisco, 1968.
6. L. H. Koopmanns, *Spectral Analysis of Time Series*, Academic Press, New York, 1974.
7. W. B. Davenport, *Probability and Random Processes*, McGraw-Hill Book Company, New York, 1970.
8. D. R. Brillinger and M. Rosenblatt, "Asymptotic Theory of Estimates of k th Order Spectra," in *Spectral Analysis of Time Series*, B. Harris, ed., John Wiley & Sons, Inc., New York, 1967.
9. P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra," *IEEE Trans. Audio Electroacoust.*, Vol. AU-15, June 1970, pp. 70-73.
10. T. G. Stockham, Jr., "High-Speed Convolution and Correlation," *Spring Joint Computer Conf., AFIPS Proc.*, Vol. 28, Spartan Books, Washington, D.C., 1966, pp. 229-233.
11. C. M. Rader, "An Improved Algorithm for High-Speed Autocorrelation with Applications to Spectral Estimation," *IEEE Trans. Audio Electroacoust.*, Vol. AU-18, Dec. 1970, pp. 439-441.

PROBLEMI

1. Sia $X(e^{j\omega})$ la trasformata di Fourier di una sequenza $x(n)$ reale di lunghezza finita, che è nulla al di fuori dell'intervallo $0 \leq n \leq N-1$. Il periodogramma $I_N(\omega)$ è definito nella formula (11.24) come la trasformata di Fourier della stima dell'autocorrelazione lunga $2N-1$ punti

$$c_{xx}(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n)x(n+m) \quad |m| \leq N-1.$$

Dimostrare che il periodogramma è legato alla trasformata di Fourier della sequenza di lunghezza finita nel modo seguente:

$$I_N(\omega) = \frac{1}{N} |X(e^{j\omega})|^2.$$

2. La stima dello spettro smussata $S_{xx}(\omega)$ è definita come

$$S_{xx}(\omega) = \sum_{m=-(M-1)}^{M-1} c_{xx}(m)w(m)e^{-j\omega m},$$

dove $w(m)$ è una sequenza finestra di lunghezza $2M-1$. Dimostrare che risulta

$$E[S_{xx}(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} E[I_N(\theta)] W(e^{j(\omega-\theta)}) d\theta,$$

dove $W(e^{j\omega})$ è la trasformata di Fourier di $w(n)$.

3. Per ricavare la relazione (11.56) abbiamo usato l'identità

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\sin [\theta N/2]}{N \sin [\theta/2]} \right)^2 d\theta = \frac{1}{N}.$$

- (a) Trovare dapprima la trasformata di Fourier della sequenza

$$s(n) = \frac{1}{N} \quad 0 \leq n \leq N-1$$

$$= 0 \quad \text{altrove.}$$

- (b) Usare il teorema di Parseval per dimostrare l'identità di sopra.

4. Nel par. 11.4.2 abbiamo studiato il metodo delle finestre usato per smussare le stime dello spettro. È stato dimostrato che il rapporto tra la varianza di una stima smussata dello spettro e la varianza del periodogramma è

$$R = \frac{\text{var}[S_{xx}(\omega)]}{\text{var}[I_N(\omega)]} = \frac{1}{N} \sum_{m=-(M-1)}^{M-1} w^2(m) = \frac{1}{2\pi N} \int_{-\pi}^{\pi} W^2(e^{j\omega}) d\omega.$$

dove N è la lunghezza della sequenza analizzata e $2M-1$ è la lunghezza totale della finestra. Quindi la varianza di $S_{xx}(\omega)$ può essere resa più piccola di quella del periodogramma scegliendo opportunamente la forma e la lunghezza della finestra.

Un'altra misura dell'effetto smussante (perdita di risoluzione) di una finestra è costituita dalla larghezza del suo lobo principale. Definiamo qui questa larghezza come l'intervallo (simmetrico rispetto all'origine) tra la prima frequenza negativa e la prima positiva per cui è $W(e^{j\omega}) = 0$.

In questo problema studieremo queste proprietà per le seguenti tre finestre:

Rettangolare

$$w_R(m) = 1 \quad |m| \leq M-1$$

$$= 0 \quad \text{altrove}$$

Bartlett

$$w_B(m) = 1 - |m|/M \quad |m| \leq M-1$$

$$= 0 \quad \text{altrove}$$

Coseno rialzato

$$w_H(m) = \alpha + \beta \cos(\pi m/(M-1)) \quad |m| \leq M-1$$

$$= 0 \quad \text{altrove.}$$

(Se $\alpha = \beta = 0.5$ questa è la finestra di Hanning e se $\alpha = 0.54$ e $\beta = 0.46$ è quella di Hamming).

- (a) Trovare la trasformata di Fourier di ciascuna di queste finestre, cioè calcolare $W_R(e^{j\omega})$, $W_B(e^{j\omega})$ e $W_H(e^{j\omega})$. Disegnare l'andamento di queste funzioni di ω .
- (b) Mostrare che per ciascuna di queste finestre sono giusti i valori che compaiono nella tabella seguente (assumere che sia $M \gg 1$).

Finestra	Larghezza del lobo principale (approssimata)	Rapporto delle varianze R (approssimato)
Rettangolare	$2\pi/M$	$2M/N$
Bartlett	$4\pi/M$	$2M/(3N)$
Coseno rialzato	$3\pi/M$	$2M(\alpha^2 + \beta^2/2)/N$

5. Nel metodo di Welch una sequenza lunga N viene suddivisa in K sottosequenze

$$x^{(i)}(n) = x(n + iM - M) \quad 0 \leq n \leq M-1, \quad 1 \leq i \leq K,$$

e a queste sottosequenze viene applicata una finestra prima di calcolare i periodogrammi modificati

$$J_M^{(i)}(\omega) = \frac{1}{MU} \left| \sum_{n=0}^{M-1} x^{(i)}(n) w(n) e^{-j\omega n} \right|^2 \quad 1 \leq i \leq K,$$

dove è

$$U = \frac{1}{M} \sum_{n=0}^{M-1} w^2(n).$$

La stima dello spettro è definita come

$$B_{xx}^w(\omega) = \frac{1}{K} \sum_{i=1}^K J_M^{(i)}(\omega).$$

Dimostrare che risulta

$$E[B_{xx}^w(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\theta) W(e^{j(\omega-\theta)}) d\theta$$

con

$$W(e^{j\omega}) = \frac{1}{MU} \left| \sum_{n=0}^{M-1} w(n) e^{-j\omega n} \right|^2$$

Suggerimento: usare il fatto che se $x(n)$ è a media nulla si ha

$$\phi_{xx}(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_{xx}(\omega) e^{j\omega m} d\omega.$$

6. Sia $x(n)$ un segnale aleatorio reale e stazionario. La stima dello spettro alla Bartlett, definita nella formula (11.47), è

$$B_{xx}(\omega) = \frac{1}{K} \sum_{i=1}^K J_M^{(i)}(\omega).$$

Consideriamo la trasformata di Fourier inversa di $B_{xx}(\omega)$, che indichiamo con $b_{xx}(m)$.

(a) Dimostrare che è

$$E[b_{xx}(m)] = \phi_{xx}(m) w_B(m)$$

con

$$w_B(m) = 1 - |m|/M \quad |m| \leq M-1$$

$$= 0 \quad \text{altrove}$$

- (b) Supponiamo ora di calcolare $B_{xx}(2\pi k/M)$ usando il metodo basato sulla FFT illustrato nel par. 11.6.1. Definiamo allora $b_{xxp}(m)$ come

$$b_{xxp}(m) = \frac{1}{M} \sum_{k=0}^{M-1} B_{xx}(2\pi k/M) e^{j(2\pi km/M)}$$

Ricavare un'espressione per $E[b_{xxp}(m)]$.

- (c) Come si modificano i risultati di (a) e (b) per il metodo di Welch? Cioè, ripetere (a) e (b) cominciando con $B_{xx}(\omega)$ espresso dalla (11.63).

